Persuasion Through Reviewers: Implementation Observability and Commitment

Mengqi Zhang*

June 9, 2026

Latest Version

(**See Appendix for Proofs**)

Abstract

We study persuasion when a sender communicates through reviewers whose private reporting types map states into signals. The sender garbles information by managing the distribution of these reviewer mappings. When the receiver observes only a coarse statistic of this distribution, the sender can secretly substitute reviewer types within an observational cell. Observability of implementation therefore becomes a commitment problem. This mapping-management perspective structurally characterizes partial commitment and its effect on persuasion. The sender can commit to an information structure only if it lies in the convex hull of the cellwise mappings selected by its own value vector. This characterization turns partial commitment into a geometric admissibility constraint on the sender's strategies. The constraint impairs persuasion by forcing the sender to choose a suboptimal admissible strategy, or by making the standard-persuasion optimum admissible only at a discounted value. For policy design, the framework indicates which reviewer types should be observable for given objectives, including robustness, rather than treating full transparency as the only benchmark.

*Media Forensics Hub and Marketing Department, Clemson University. Email: mengqi@clemson.edu. I bear complete responsibility for all the errors present in the paper.

1 Introduction

In the conventional Bayesian persuasion framework, the sender (she/her) designs an information structure as a mapping from states into signals delivered to the receiver (he/him). This abstraction makes the persuasion problem tractable but leaves open how the information structure is implemented in many real-world persuasion practices. Firms rely on testimonials and rating systems. Political campaigns and advocacy groups rely on supporters, commentators, and local messengers. To make persuasion less costly, the sender reduces direct interaction with receivers; to make the signal more credible, the sender delegates information transmission to third parties. In these cases, the sender does not speak to the receiver directly. Persuasion is implemented through a population of reviewers.

This paper studies persuasion through such reviewers. Depending on their background, experience, and preferences, reviewers may interpret the same objective evidence or product quality in different ways. As a reduced-form persuasion technology, we model heterogeneous reviewers as non-strategic reporting types. Technically, each reviewer type is a deterministic mapping from states to signals. For the receiver who randomly encounters reviewers,[1] these mappings are rarely transparent. A consumer reading a product review or a Reddit comment usually does not know the reviewer's full history, private preference, ideology, or incentive that determines the reporting rule. Even platforms that label some reviews as "verified" do not fully reveal the reviewer's information or reporting type. By changing the composition of reviewer types, the sender can influence the receiver's belief about the mapping rule associated with the signal he observes. When the receiver fully observes this distribution, the sender can replicate the outcome of any standard persuasion problem in this indirect persuasion environment, as long as the reviewer type space is rich enough (Lemma 1).

[1]The receiver may read one review, a few reviews, or an aggregate rating. In the model, we consider the baseline case where the receiver randomly meets one reviewer. Allowing the receiver to observe several reviews would enlarge the signal space, but it would not remove the underlying issue of partial observability and the mechanism of persuading through reviewers under this friction.

However, this theoretical equivalence is only a benchmark. The central difficulty is that the receiver usually observes only coarse information about the review process, not the full reviewer distribution. Some platforms, such as TripAdvisor, Yelp, and Google, have historically allowed broad participation in posting reviews, making it difficult for users to infer the reviewer base. Other platforms, such as Expedia and Amazon, restrict reviews to verified transactions or attach verified-purchase labels. Such verification refines the observable reviewer types and may exclude some fake reviews, but does not identify the preferences or reporting rules of genuine reviewers. A verified purchaser can be optimistic or skeptical, experienced or inexperienced, ideologically aligned, or hostile. These differences can change how the same product quality is translated into a signal. In this case, the observable statistic remains coarse.

The sender's ability to secretly manage the reviewer distribution then creates a commitment problem. If the receiver observes only a coarse statistic of this distribution, the sender can secretly adjust the reviewer composition within an unobserved pool. Fixing the receiver's action rule, the sender would like to shift mass toward reviewer types that send signals favorable to her objective. Delegating signal delivery to reviewers can prevent the sender from withholding a realized signal, but it does not make her commit to the distribution of technologies that generate signals. Review manipulation is empirically documented in online platforms (Mayzlin et al., 2014; Luca and Zervas, 2016). At the same time, review systems remain effectively persuasive and economically important, as empirically supported by online book sales and restaurant revenue (Chevalier and Mayzlin, 2006; Luca, 2016). How can review systems remain persuasive even when reviewer composition is partially observable and thus manipulable? More structurally, how does partial observability of the reviewer distribution shape persuasion?

Our paper contributes by making this commitment problem structural to characterize how it shapes standard persuasion, where the sender's payoff is state independent. Abstract information-structure representation is mathematically useful in persuasion problems but

comes at the cost of its connection to how messages are generated and conveyed in practice. This makes it difficult not only to justify the commitment assumption (Kamenica et al., 2021; Deb et al., 2026), but also to analyze how relaxing that assumption changes persuasion. By capturing the specific but common scenario of persuading through reviewers, we change the perspective of standard persuasion from *mapping design* to *mapping management*. The sender does not directly design a stochastic mapping. She manages a distribution over deterministic mappings that represent reviewer types. An observability structure determines which mappings the receiver can distinguish and which mappings are pooled. Conditional on the observable statistic, the sender faces a fiber of reviewer distributions within which deviations are hidden. Imperfect commitment therefore becomes an endogenous constraint on the feasible set of effective information structures.

Partial observability modifies the purely sequential structure of a standard persuasion game. We use a reduced representation that separates the sender's choice into an observable component and a hidden within-fiber component. First, the sender chooses the observable statistic of the reviewer distribution. This statistic defines a fiber of reviewer distributions that remain hidden from the receiver. Second, conditional on this statistic, the sender's hidden within-fiber distribution and the receiver's action rule are chosen simultaneously in the reduced continuation game. This perspective is related to Lin and Liu (2024), who study credible persuasion when the receiver observes the final message distribution. Because the coarse statistic that the receiver observes need not pin down the final signal distribution, a commitment problem persists even when the sender's payoff is state independent.

In the static continuation game, the sender's problem is a linear program over a polyhedral constraint. Pure-strategy equilibria select extreme points or faces of the fiber, where dominated mappings are eliminated from the support of the reviewer distribution (Lemma 2). This endogenous mapping selection yields a tractable geometry (Proposition 1). We define the concept of ***admissibility*** through equilibrium in this continuation game. An information structure is admissible if and only if it lies in the convex hull generated by

the reviewer mappings that its own induced value vector selects as sender-optimal within each observational cell (Theorems 1 and 2). Thus, the sender's value function plays two roles. It is the objective in the persuasion problem and it also selects the constraint set that determines whether the intended information structure can be implemented. A mismatch between these two roles invalidates the sender's commitment to the information structure.

This framework nests several benchmarks. Full observability of the reviewer distribution collapses each fiber to a singleton and restores the standard persuasion benchmark (Corollary 2). Message-distribution observability, as in Lin and Liu (2024), which can be reproduced by extending the observability statistic beyond the baseline partition of reviewer mappings, is another case where full commitment applies under the sender's state-independent payoff. When the observable statistic is the final signal distribution, the sender's state-independent payoff is constant on the fiber. Since this fiber coincides with the sender's indifference set, credibility is nonrestrictive. At the opposite end, uninformative communication (cheap talk) is robust regarding admissibility as long as it is within the fiber, because all signals induce the same posterior belief and the sender has no incentive to change their composition. Between these cases, partial observability creates a nontrivial admissible region. In the binary case, the characterization reduces to a simple wedge condition, showing exactly when the full-commitment benchmark is ruled out and how the sender optimally adjusts (Proposition 2).

We treat mixed-strategy equilibria as a residual source of admissibility. Whenever a pure-strategy equilibrium exists on a fiber, ***pure admissibility*** is the only relevant implementation notion. If there is no pure-strategy equilibrium, a receiver mixture can neutralize the sender's incentive to move among hidden mappings (Theorem 3). This enlarges the selected face of the fiber (Corollary 1), giving rise to ***expected admissibility***. However, the remedy comes with a price. The receiver's strategy mixture that neutralizes the sender's hidden deviation generally discounts the sender's expected payoff (Corollary 3). If pure admissibility rules out the full-commitment benchmark, the information cost is unavoidable.

Even if the receiver's mixed strategy can restore expected admissibility for the benchmark information structure, it does not restore the benchmark persuasion effectiveness.

An important implication is that imperfect commitment need not cause persuasion to collapse. Different observability structures generate different commitment constraints and, therefore, different persuasion outcomes. This helps reconcile the motivating puzzle and explain why review and rating systems can be both manipulable and effective. The policy implication is also different from a simple "more disclosure is always better" principle. Disclosing additional reviewer mappings can be costly and may not strictly improve the sender's expected payoff. What matters is not only how many reviewer categories are observed, but which mappings are identifiable. A policy designer should identify which pooled mappings threaten admissibility and which pools can be safely tolerated before refining the observability structure.

We adopt this spirit to study the policy design for robust admissibility. A platform, regulator, or certifier often influences the observability structure before knowing the particular persuasion problem that will arise. The policy question is therefore which observability structures make admissibility robust across different persuasion problems. Full observability is trivially sufficient for this robustness, but often impossible or too costly. We show that a proper subset of reviewer types can be sufficient to recover full admissibility up to signal relabeling (Theorem 4). In the binary case, even coarser observability can be enough when either the target experiment or the sender's value ranking is known (Proposition 3). We also extend this robustness discussion to the persuasion stage. In standard persuasion, once the experiment is known, the sender's prior is irrelevant to the receiver. However, under partial observability, the sender's prior helps the receiver infer which hidden mapping the sender prefers in the static continuation game. We characterize observability structures that are robust to private sender priors (Theorem 5).

The rest of the paper proceeds as follows. Section 2 discusses related literature. Sections 3 and 4 introduce the model and the deterministic-mapping representation of information

structures as the analytical framework. Section 5 characterizes pure admissibility based on pure-strategy equilibria in the continuation game and studies its effect on persuasion outcomes. Section 6 introduces expected admissibility through mixed-strategy equilibria in the continuation game. Section 7 analyzes robust observability structures and policy design. Section 8 concludes.

2 Related Literature

This paper is closest to Lin and Liu (2024) and Deb et al. (2026). Lin and Liu introduce credible persuasion when the receiver observes the final distribution of messages but not the full information structure. Our paper studies a different observability primitive. The receiver observes a coarse statistic of the implementation technology, but not necessarily the final signal distribution. This distinction matters most sharply when the sender's payoff is state independent. With this state-independent payoff, our framework can mathematically accommodate Lin and Liu's model, where the observable statistic is specialized to the final message distribution, and the fiber coincides with the sender's indifference set.[2] The non-trivial commitment problem appears only away from that special observability statistic. The distinction in observable primitives, therefore, leads to different findings and complementary applications.

Deb et al. (2026) study an indirect persuasion environment where an uninformed principal persuades a receiver through a strategic informed agent. They reinterpret the commitment to information structures as the commitment to employment contracts and show that, under their separability assumptions, commitment to contracts can implement the same outcome as the full-commitment persuasion benchmark. Our paper essentially takes the opposite perspective. Rather than justifying full commitment, we take the implementation technologies and their observability structure as primitives and ask which information

[2]See Corollary 4 in the Appendix for detailed discussions.

structures remain admissible under partial commitment.

A broader literature studies the commitment assumption in persuasion. Early optimal disclosure and persuasion models, including Rayo and Segal (2010) and Kamenica and Gentzkow (2011), take commitment to an information structure as the key exogenously assumed primitive. Subsequent work asks when this assumption can be justified, discharged, or reinterpreted. Lipnowski and Ravid (2020) show that informative communication can arise without commitment when the sender degrades self-serving information. Kamenica and Lin (2025) relate the value of commitment to the value of randomization. Verifiable-information models, including Titova and Zhang (2025), Arieli and Stewart (2025), and Shishkin et al. (2026), examine when hard evidence or uncertainty about the sender's evidence can substitute for commitment. Dynamic and reputational approaches, including Margaria and Smolin (2018), Fudenberg et al. (2022), Pei (2023), Best and Quigley (2024), Mathevet et al. (2024), and Kuvalekar et al. (2022), study how repeated interaction, reputation, or relational incentives can support commitment-like payoffs.

Another branch imposes partial commitment through explicit frictions. Papers in this branch can be organized along two dimensions: whether the friction is external or intrinsic to the information process, and whether the commitment problem appears at signal generation or signal delivery. Min (2021) and Lipnowski et al. (2022) assume that commitment may fail with an exogenous probability. Guo and Shmaya (2021) model the commitment power through the cost of forecast miscalibration. Zhou (2023) studies costly verification when the sender can choose both an exposed and a hidden communication plan, and the receiver can pay to verify which plan is used. Nguyen and Tan (2021), Ederer and Min (2025), and Kreutzkamp and Lou (2025) study ex post signal-delivery problems in which the sender can misreport a realized signal subject to costs, detection, or credibility constraints. Perez-Richet and Skreta (2022) allow falsification of test inputs, moving the friction to the signal-generation side. Jiang (2024) studies a private sequential evidence acquisition, where the sender cannot commit to how much evidence she acquires or fully disclose the acquired

signal history. Lou (2023), Dai et al. (2026), and Antic and Pei (2026) study selective disclosure and data truncation, where the sender can hide or selectively reveal realized evidence. We emphasize the environment where the sender can secretly change the ex ante composition of the reviewer mappings that generate signals. Partial commitment is therefore an intrinsic moral-hazard problem created by implementation observability at the signal generation stage. Compared to previous studies in this branch, our paper occupies a cell in the two-dimensional classification that differs from each of them in at least one dimension.

The mechanism of our paper is related to three topics in persuasion: hidden dimensions of information, indirect persuasion, and coarse communication. Libgober (2022) analyzes multidimensional experiments in which some dimensions of the experimental process are hidden, while Celik and Drugov (2025) study score disclosure when aggregate scores rather than full attributes are revealed. They agree that the observability of the information-generating process matters, but ask different questions than we do. Papers on mediated or indirect persuasion, including Kosenko (2020), Bizzotto et al. (2022), Arieli et al. (2022), Corrao and Dai (2024), and Lagziel and Lehrer (2025), typically feature strategic intermediaries and mediators whose incentives directly constrain communication. In contrast, our reviewers are nonstrategic mapping types. This removes the intermediary-incentive problem and isolates a different source of commitment failure connected to an endogenous coarseness constraint on mappings. Finally, papers on constrained or coarse persuasion often impose coarseness directly on the mapping from states to messages through the message space, channel, or communication technology (Le Treust and Tomala, 2019; Gradwohl et al., 2022; Aybas and Turkel, 2025; Lyu et al., 2025). In our model, coarseness is endogenously induced by the partial observability of reviewer mappings. It is therefore structural rather than purely cardinal. Different observability partitions with the same number of cells can generate different sets of admissible information structures, depending on which mappings are pooled and on the persuasion problem.

Our policy discussion connects to robust persuasion. Robust persuasion papers typically ask which information structures protect the sender against the worst-case primitive in the game, including the receiver's belief, utility, and private information (Hu and Weng, 2018; Dworczak and Pavan, 2022; Kosterina, 2022; Babichenko et al., 2022). Our robustness inquiry is different. We ask which observability structures make admissibility robust as the persuasion problem varies. Furthermore, we consider the sender's prior belief as an unknown primitive for the receiver. This discussion arises only when robustness concerns meet partial commitment, where the receiver must infer which hidden mapping the sender would prefer based on her prior belief. Also related to our policy discussion, Vong (2025) shows in a dynamic reputation model that full transparency of output histories can undermine efficiency, while coarse ratings can virtually implement efficient effort. This complements our policy message that full observability is not always the right benchmark.

3 Model

As in conventional direct Bayesian persuasion problems, in this model there is a sender (she, her) and a receiver (he, him). There is a finite set of possible states $\omega \in \Omega$. The receiver chooses the action $a \in \mathcal{A}$, which determines his payoff, $u : \Omega \times \mathcal{A} \to \mathbb{R}$, together with the actual state. This action solely determines the sender's payoff $v : \mathcal{A} \to \mathbb{R}$. The sender and the receiver have prior beliefs in the actual state $p_s \in \Delta(\Omega)$ and $p_r \in \text{int}(\Delta(\Omega))$, respectively. Since the receiver is uncertain about the actual state, his optimal action is based on his belief. Let $\mathcal{S} = \{s_1, \ldots, s_n\}$ be a finite signal set. The sender wants to send signals using an information structure $\Pi = (\pi(\cdot \mid \omega))_{\omega \in \Omega} \in \Delta(\mathcal{S})^{\Omega}$, where $\pi(s \mid \omega)$ is the probability of signal s in state ω. By Bayes' rule, the realized signal can induce the receiver to choose the action that the sender desires. However, the signal needs to be sent through reviewers.

Reviewers, with total mass normalized to 1, are informed of the true state. They may

have different opinions on the same issue. For example, different customers may have different preferences about the color of an iPhone. Reflecting this fact, reviewers in our model non-strategically and deterministically interpret the state they observe as different signals. Let Θ denote the reviewer-type set. $\gamma : \Theta \times \Omega \to \mathcal{S}$ captures this interpretation. Each reviewer type θ induces a deterministic mapping $\Omega \to \mathcal{S}$ summarized by matrix Π_θ, where $(\Pi_\theta)_{ij} = \mathbf{1}\{s_i = \gamma(\theta, \omega_j)\}$. A reviewer distribution $\lambda \in \Delta(\Theta)$ induces the effective experiment $\Pi(\lambda) := \sum_{\theta\in\Theta} \lambda(\theta)\Pi_\theta$. Since these mappings representing the reviewer type are deterministic, there are no more than n^n types of reviewer in a problem with $|\Omega| = n$ and $|\mathcal{S}| = n$.

The sender designs the distribution over reviewer types $\lambda \in \Delta(\Theta)$. Once the choice is made, a reviewer is randomly drawn according to this distribution, meets the receiver, and passes the signal to him. The receiver does not observe the type of reviewer that delivers the signal, but can observe certain statistics of the reviewer distribution, $z = T\lambda$, where $T \in \{0, 1\}^{B\times|\Theta|}$ and $\mathbf{1}_B^\top T = \mathbf{1}_{|\Theta|}^\top$. The structure of the matrix T imposes a partition of the reviewer-type set Θ into observational cells, $\Theta_k := \{\theta \in \Theta : T_{k\theta} = 1\}, k = 1, \ldots, B$, so the receiver only observes $z_k = \sum_{\theta\in\Theta_k} \lambda(\theta)$, $k = 1, \ldots, B$. Once the receiver receives the signal, he updates his prior belief to his posterior belief $q_r \in \Delta(\Omega)$, based on his observation of z and Bayes' rule. All primitives and rules in the game, unless specified, are common knowledge.

The persuasion game starts with Nature randomly choosing the true state $\omega^* \in \Omega$ according to the objective state draw $p_0 \in \Delta(\Omega)$ and informs the reviewers privately. The beliefs p_s and p_r are the sender's and receiver's subjective beliefs about this state. They may differ from p_0. The sender then chooses the optimal reviewer distribution λ^* based on her belief p_s,

$$\lambda^* \in \arg\max_{\lambda\in\Delta(\Theta)} \sum_{\theta\in\Theta}\sum_{\omega\in\Omega} \lambda(\theta)p_s(\omega)v\Big(a^*\big(\gamma(\theta,\omega), T\lambda\big)\Big) \tag{1}$$

in anticipation of the receiver's best response $a^*(\cdot)$ to the signal he receives from the reviewer. After the sender chooses λ, the receiver observes $z = T\lambda$; then a reviewer $\theta \in \Theta$ is randomly drawn according to λ, after which the receiver observes s delivered by this reviewer.

Upon observing the statistic z about the sender's choice of reviewer distribution λ, the receiver infers (by equilibrium reasoning) that the sender chose the (generically unique) within-fiber optimizer,

$$\lambda^*_{(z)} \in \arg \max_{\lambda \in \Delta(\Theta): T\lambda = z} \sum_{\theta \in \Theta} \sum_{\omega \in \Omega} \lambda(\theta) p_s(\omega) v\Big(a^*\big(\gamma(\theta, \omega), z\big)\Big). \tag{2}$$

Based on this information about $\lambda^*_{(z)}$ and the signal s he observes, the receiver updates his belief about the state by Bayes' rule and chooses his action $a^*(s, z)$,

$$a^*(s, z) \in \arg \max_{a \in \mathcal{A}} \sum_{\omega \in \Omega} q_r(\omega \mid s, z) u(\omega, a), \tag{3}$$

where

$$q_r(\omega | s, z) = \frac{\sum_{\theta \in \Theta} \mathbf{1}\big[s = \gamma(\theta, \omega)\big] p_r(\omega) \lambda^*_{(z)}(\theta)}{\sum_{\omega' \in \Omega} \sum_{\theta \in \Theta} \mathbf{1}\big[s = \gamma(\theta, \omega')\big] p_r(\omega') \lambda^*_{(z)}(\theta)}.$$

When a pure receiver best response must be selected from a non-singleton best-response set, we use a fixed sender-favorable selection, with arbitrary but fixed tie-breaking among sender-equivalent actions. The mixed-admissibility section separately allows the receiver to randomize over best-response action rules.[3] Equivalently, $a^*(\cdot, z)$ is an action rule in $\mathcal{A}^{\mathcal{S}}$. In later sections, we denote generic action rules by α.

Finally, the payoffs for both the sender and the receiver are realized on the basis of

[3]To avoid zero-probability signals and ensure that the denominator of $q_r(\omega|s, z)$ is always positive, assume that after choosing the intended signal $\gamma(\theta, \omega)$, a reviewer trembles with an $\epsilon > 0$ chance and instead sends a signal drawn from a full-support distribution within $\text{int}(\Delta(\mathcal{S}))$. We assume that $\epsilon \to 0$ and define posterior using Bayes' rule in the $\epsilon-$perturbed game. Since payoffs are bounded, we can understand that this perturbation does not affect the sender's objective.

the true state and the receiver's action. Since the receiver observes $z = T\lambda$ but not λ, the original extensive form has imperfect information. After observing z, the receiver's information set contains all reviewer distributions λ such that $T\lambda = z$.[4] Rather than working directly with beliefs at every information set, we use a reduced representation that separates the observable and hidden components of the sender's choice. The sender first chooses the observable statistic z. Conditional on z, the sender chooses a hidden reviewer distribution satisfying $T\lambda = z$, while the receiver chooses an action rule based on the information structure he expects. This continuation interaction is a *de facto* static game induced by z. We analyze subgame-perfect equilibrium of the reduced z-stage game, where Nash equilibrium is the relevant solution concept for each fiber continuation game. The next section defines this continuation game formally and uses it to define admissibility.

4 Framework

The direct Bayesian persuasion problem has been well studied. In this paper, we modify the signal-delivery mechanism to give a concrete microfoundation for the signal-generating process. In particular, randomization that is implemented directly through an information structure $\Pi \in \Delta(\mathcal{S})^{\Omega}$ in standard Bayesian persuasion is implemented here through the sender's design of a reviewer distribution $\lambda \in \Delta(\Theta)$ based on reviewers' deterministic interpretations $\gamma : \Theta \times \Omega \to \mathcal{S}$. If, for a given reviewer set, every direct information structure can be generated by some reviewer distribution, then persuasion through reviewers and direct persuasion are equivalent. This equivalence gives a clean benchmark framework. It lets us attribute key abstract concepts in persuasion, such as information structures and the commitment assumption, to reviewer availability and observability and isolate which constraints on them are responsible for departures from the standard direct-persuasion frontier.

[4] In the original extensive form, the natural background solution concept is perfect Bayesian equilibrium, because the receiver must form beliefs over hidden reviewer distributions after observing z.

Fix $|\Omega| = |\mathcal{S}| = n < \infty$. Let

$$\mathcal{P} \equiv \Delta(\mathcal{S})^{\Omega} \cong \left\{ \Pi \in \mathbb{R}_{+}^{n \times n} : \mathbf{1}_n^{\top} \Pi = \mathbf{1}_n^{\top} \right\}$$

denote the set of all $n \times n$ column-stochastic matrices Π, where rows are indexed by signals and columns by states, so $\Pi_{ij} = \pi(s_i|\omega_j)$.

Lemma 1 (Benchmark framework)**.** *Suppose Θ contains all deterministic mappings $\gamma(\theta, \cdot) : \Omega \to \mathcal{S}$ and so $|\Theta| = n^n$. Then for every information structure $\Pi \in \mathcal{P}$ there exists $\lambda \in \Delta(\Theta)$ such that for all $s \in \mathcal{S}$ and $\omega \in \Omega$,*

$$\pi(s \mid \omega) = \sum_{\theta \in \Theta} \lambda(\theta) \mathbf{1}\{s = \gamma(\theta, \omega)\}.$$

According to Kamenica and Gentzkow (2011), with full commitment, the optimal direct information structure uses at most n signals when there are n states. This justifies our baseline n-by-n $\Pi-$space $\mathcal{P}$, which is rich enough for frictionless direct persuasion, where full commitment is assumed and all information structures in this space are feasible. Lemma 1 implies that $\mathcal{P}$ is a polytope whose extreme points are exactly the n^n deterministic response matrices $\{0,1\}^{n \times n}$ induced by reviewer types. This formalizes each effective information structure as a distribution over a subset of these n^n reviewer types. Throughout the paper, we take this "full reviewer space" as the baseline framework and study how the reviewer base and the observability structure restrict the attainable set.

A key difference between direct persuasion and indirect persuasion through reviewers is signal labeling. In persuasion through reviewers, signal labels are exogenously pinned down by the fixed reviewer base. Each type θ implements a fixed mapping $\gamma(\theta, \cdot)$, connecting the state to these exogenous labels. As a result, different Π are different objects in $\mathcal{P}$ once constraints on reviewer availability and observability are imposed. In contrast, from

the perspective of direct persuasion, where only persuasion outcomes and the associated posteriors matter, these labels are merely names because row permutations of Π are outcome-equivalent.

Because labels are fixed in indirect persuasion, the full reviewer space $\mathcal{P}$ contains many isomorphic copies of the same persuasion outcomes. To connect indirect persuasion back to the direct benchmark, it is useful to group these isomorphic copies. Formally, let Σ be the set of permutations of $\{1, \ldots, n\}$ and define the relabeling operator $(\sigma\Pi)_{ij} \equiv \Pi_{\sigma^{-1}(i),j}$. Then Π and $\sigma\Pi$ generate the same persuasion outcome. This analysis approach motivates the focus on a subspace of $\mathcal{P}$ formed by representatives from each relabeling class.

Definition 1. *A set $\mathcal{D} \subseteq \mathcal{P}$ is a **fundamental region** if for every $\Pi \in \mathcal{P}$ there exist a permutation $\sigma \in \Sigma$ and $\widetilde{\Pi} \in \mathcal{D}$ such that $\Pi = \sigma\widetilde{\Pi}$. It is **proper** if, for $\Pi, \Pi' \in \mathcal{D}$, $\Pi = \sigma\Pi'$ implies $\Pi = \Pi'$, up to measure-zero tie cases.*

Proper fundamental regions are defined by how signals are labeled based on certain orderings. In alignment with the persuasion problem, the economically meaningful order is about how signals move the receiver's belief. Consistent with $\Pi = \{\pi(s_i|\omega_j)\}$, we index signals by $i \in \{1, \ldots, n\}$ and states by $j \in \{1, \ldots, n\}$. Fix a strictly increasing scalar statistic $m(\omega_j)$. Without loss of generality, let $m(\omega_j) = j$. Given the receiver's prior belief $p_r \in \text{int}(\Delta(\Omega))$, define the scalar posterior mean associated with the signal s_i in experiment Π as

$$\bar{p}(s_i; \Pi) \equiv \sum_{j=1}^{n} j \cdot \frac{\pi(s_i \mid \omega_j) p_r(\omega_j)}{\sum_{\ell=1}^{n} \pi(s_i \mid \omega_\ell) p_r(\omega_\ell)}.$$

Definition 2. *For a permutation $\sigma \in \Sigma$, a **canonical region** $\mathcal{C}_\sigma$ is a set of $\Pi \in \mathcal{P}$ such that $\bar{p}(s_{\sigma(1)}; \Pi) \geq \bar{p}(s_{\sigma(2)}; \Pi) \geq \cdots \geq \bar{p}(s_{\sigma(n)}; \Pi)$, with strict inequalities holding in the interior and ties occurring only on measure-zero boundaries. Each $\mathcal{C}_\sigma$ is a proper fundamental region up to the boundaries.*

Consider a binary case where $n = 2$, characterizing whether a product potentially has a feature (ω_1) or not (ω_2), and two signals represent recommendations to buy it (s_1) or not (s_2). The sender can fully manipulate the effective information structure if there are four reviewer types defined by four deterministic maps $\gamma(\theta_k, \cdot)$ and, specifically,

$$\begin{aligned}
\theta_1 &: \ \omega_1 \to s_1, \quad \omega_2 \to s_2 \\
\theta_2 &: \ \omega_1 \to s_2, \quad \omega_2 \to s_1 \\
\theta_3 &: \ \omega_1 \to s_1, \quad \omega_2 \to s_1 \\
\theta_4 &: \ \omega_1 \to s_2, \quad \omega_2 \to s_2.
\end{aligned}$$

Reviewers $\{\theta_1, \theta_2\}$ can be interpreted as previous customers who have different opinions on this feature of the product: one thinks that it is a fantastic design, but the other believes that it is a redundancy that causes the price to be unnecessarily high. $\{\theta_3, \theta_4\}$ are just those reviewers who are "yes-men" saying good to everything and who are picky enough to leave bad reviews for the product regardless. These types capture common heterogeneity in real-world review populations. With this demography, the seller who launches a new product, for example, can easily manipulate the trial review if she can fully manage whom her samples are distributed to during the testing phase.[5]

To represent the full reviewer space, construct a two-dimensional square space with horizontal axis $\pi(s_2|\omega_1) \in [0, 1]$ and vertical axis $\pi(s_2|\omega_2) \in [0, 1]$. Canonical regions, defined as two halves demarcated by the 45-degree line, are naturally fundamental regions. Swapping the two signal labels (equivalently, permuting the two rows of Π) sends the point $(\pi(s_2|\omega_1), \pi(s_2|\omega_2)) = (\pi_1, \pi_2)$ above the 45-degree line to $(\pi(s_2|\omega_1), \pi(s_2|\omega_2)) = (1 - \pi_1, 1 - \pi_2)$ below the 45-degree line, and this relabeling preserves the persuasion outcome. This equivalence can matter in indirect persuasion when the reviewer base makes one labeling feasible but the other infeasible. For example, if there are many reviewers who, for some reason, such as to preserve exclusivity, do not recommend a product when it has a good

[5]This practice can continue even into the regular product selling phase, as long as the seller can decide who her product is sold to, but this may incur costs in sales due to customer selection.

feature, but those reviewers who recommend when observing this feature are limited, the sender can instead use "not recommending" as a signal of a good feature. As long as consumers can observe the reviewer distribution, this substitution has the same effect on the persuasion outcome.

As shown in the binary example, the full reviewer space defined in Lemma 1 contains many isomorphic spaces that are redundant for persuasion outcomes. We nevertheless need this full space as the baseline framework, because the spaces we later use as canvases for analysis, especially those pinned down by given constraints, must be supported by a fixed reviewer base, and some of them cannot be embedded inside any smaller polytope with fewer than the full n^n deterministic vertices. In the binary case, a proper fundamental region can be any half of the square that selects one representative from each point-reflection pair about $(1/2, 1/2)$, but a fixed three-type reviewer base cannot support a space covering an arbitrary proper fundamental region.[6]

5 Partial Observability and Pure Admissibility

In persuasion through reviewers, the sender's problem is characterized by the set of attainable values and the set of reviewer distributions that attain them. The former is exogenously determined by primitives. The latter depends endogenously on equilibrium when the reviewer distribution is only partially observable to the receiver. With partial observability, the receiver does not observe λ directly, but only the coarse statistic $z = T\lambda$. Hence, two reviewer distributions that generate the same z are observationally equivalent from the receiver's perspective even if they induce different effective information structures.

For a fixed z, the sender can still reallocate the reviewer distribution within the fiber $\mathcal{P}(z) := \{\Pi(\lambda) : \lambda \in \Lambda(z)\}$, where $\Lambda(z) := \{\lambda \in \Delta(\Theta) : T\lambda = z\}$, without changing any primitive that the receiver observes directly. For this reason, the receiver's partial observ-

[6]For example, take a fundamental region that equals a canonical region with one point removed, together with the point-reflection of the removed point.

ability of the reviewer distribution changes the pure sequential structure of the standard persuasion game. After the sender chooses the observable statistic z, the continuation problem is best understood as a *de facto* static game. In this game, the sender chooses the actual and latent distribution $\lambda \in \Lambda(z)$, while the receiver chooses an action rule $a(\cdot)$ that is contingent on the realized signals and his inference from the induced information structure Π. In equilibrium, the *de facto* simultaneous moves λ and $a(\cdot)$ are mutually dependent according to (2) and (3).

In the continuation-game representation, the receiver's strategy is an action rule, which specifies an action after each possible signal. Therefore, the sender's payoff induced by the receiver's best response to an expected information structure Π is summarized by an n-dimensional value vector, with one entry for each signal. Under the pure-selection convention in the model, when the receiver has multiple best responses after a signal, we use a fixed sender-favorable selection, with arbitrary but fixed tie-breaking among sender-equivalent actions. This selection defines an action rule $\alpha^{\Pi} : \mathcal{S} \to \mathcal{A}$ conditional on the expected information structure Π, and therefore a selected sender value vector associated with Π.

Because we work within canonical regions, we write this vector in canonical signal order. If $\Pi \in \mathcal{C}_\sigma$, define $\mathbf{v}(\Pi) = (v_1(\Pi), \ldots, v_n(\Pi))^\top$ by $v_i(\Pi) = v\big(\alpha^{\Pi}(s_{\sigma(i)})\big), i = 1, \ldots, n$. Thus, the value vector attached to the original signal labels is $\sigma\mathbf{v}(\Pi)$, where $(\sigma\mathbf{v})_j = \mathbf{v}_{\sigma^{-1}(j)}$. When the candidate information structure is fixed, we write $\mathbf{v}$ for $\mathbf{v}(\Pi)$. Conditional on this value vector, if $\Pi = \sum_{\theta\in\Theta} \lambda(\theta)\Pi_\theta$ is located in the canonical region $\mathcal{C}_\sigma$, the sender's within-fiber best-response is

$$\arg \max_{\lambda' \in \Lambda(z)} (\sigma\mathbf{v})^\top \left(\sum_{\theta\in\Theta} \lambda'(\theta)\Pi_\theta \right) p_s. \tag{4}$$

When the persuasion message is delivered by the reviewer, the commitment problem arises only when the sender secretly deviates and uses another information structure to generate signals. In this sense, the pure strategy equilibrium in the *de facto* static continuation game precisely connects the sender's commitment to her incentive compatibility. We use

this connection to define which information structure is admissible in a persuasion problem.

Definition 3 (Pure admissibility)**.** *Fix an observability matrix T. An effective information structure Π is* ***purely admissible*** *if there exist a feasible observable statistic z, a reviewer distribution $\lambda^* \in \Lambda(z)$, and an action rule $\alpha^* \in \mathcal{A}^{\mathcal{S}}$ such that (λ^*, α^*) is a pure-strategy equilibrium of the* de facto *static continuation game on the fiber $\Lambda(z)$, and $\Pi = \Pi(\lambda^*) = \sum_{\theta\in\Theta} \lambda^*(\theta)\Pi_\theta$. Accordingly, a direct-persuasion outcome, summarized by the induced posteriors and the value vector $\mathbf{v}$, is* ***purely admissible*** *in persuasion through reviewers if it is induced by a purely admissible effective information structure.*

Throughout this section, admissibility is defined in pure strategies. In persuasion, the receiver cares about the ex post implementation of an effective information structure. A mixed-strategy equilibrium may match the receiver's inference and the sender's actual choice of information structure in expectation, while the mismatch can still occur in realization. When a pure-strategy equilibrium exists on the fiber, we take it as the primary implementation notion. The next section studies the residual case in which no pure-strategy equilibrium exists in a fiber, so that the receiver's strategy mixture becomes relevant. Even in that case, pure admissibility still indicates whether the persuasion problem is affected by commitment issues.

Lemma 2. *Let (λ^*, α^*) be a pure-strategy equilibrium of the* de facto *static continuation game on the fiber $\Lambda(z)$. Suppose $\Pi(\lambda^*) \in \mathcal{C}_\sigma$, and let $\mathbf{v}$ be the sender value vector induced by α^*. For each observational cell Θ_k, $k = 1, \ldots, B$, if $\theta', \theta'' \in \Theta_k$ satisfy both $\lambda^*(\theta') > 0$ and $\lambda^*(\theta'') > 0$, then $(\sigma\mathbf{v})^\top \Pi_{\theta'} p_s = (\sigma\mathbf{v})^\top \Pi_{\theta''} p_s = \max_{\theta\in\Theta_k} (\sigma\mathbf{v})^\top \Pi_\theta p_s$. Hence, whenever the maximizer within each observational cell is unique, each cell assigns positive mass to at most one reviewer type in a pure-strategy equilibrium.*

For a fixed value vector, any reviewer type that is strictly dominated within its observational cell cannot receive positive mass in a pure-strategy equilibrium. Lemma 2 therefore connects the sender's hidden-deviation incentive to the elimination of reviewer types. What remains in each observational cell is the face generated by the sender-maximizing types in that cell.[7] In this sense, Lemma 2 connects the commitment problem in persuasion through reviewers to the constraint on reviewer availability. If a reviewer type has its contribution to the sender's objective that is strictly dominated by another type within the same observational cell, this type of reviewer and the corresponding mapping can be understood as unavailable to the sender. This constraint can reduce the available mappings from n^n baseline to B in the worst case under unique within-cell maximizers, and creates coarseness in persuasion. But in contrast to many discussions of coarse persuasion, Lemma 2 emphasizes that this reduction changes not only the cardinality, but also the structure of coarseness.

Since the value vector is arbitrarily given, Lemma 2 does not yet characterize admissibility. However, it is a key step that characterizes the support of the reviewer distribution with regard to the relevance to an admissible information structure. In equilibrium, the receiver correctly infers the reviewer distribution and chooses the action to connect this reviewer distribution to the value vector. Once the value vector is determined in this way, Lemma 2 becomes a basis for establishing a geometric principle for testing admissibility.

Theorem 1. *Fix a persuasion problem and an observability matrix T with an observability structure $\{\Theta_k\}_{k=1}^{B}$. An effective information structure Π, located in the canonical region $\mathcal{C}_\sigma$ and inducing the selected value vector $\mathbf{v} = \mathbf{v}(\Pi)$, is purely admissible if and only if there exist weights $z = (z_1, \ldots, z_B) \in \Delta(\{1, \ldots, B\})$ and a set selection $M_k^\sigma(\mathbf{v}) :=$*

[7] Pure strategy refers to the sender's deterministic choice of reviewer distribution λ. It does not require singleton support within an observational cell when the sender is indifferent across the types in that cell. The support is a face in general, and the singleton is its special case.

$\arg\max_{\theta\in\Theta_k}(\sigma\mathbf{v})^\top\Pi_\theta p_s$ *for each cell* k *such that*

$$\Pi = \sum_{k=1}^{B} z_k \widehat{\Pi}_k,$$

where $\widehat{\Pi}_k \in \text{co}\{\Pi_\theta : \theta \in M_k^\sigma(\mathbf{v})\}$*, meaning that it belongs to the convex hull generated by representative faces of each observational cell selected by the value vector it induces.*

Theorem 1 formalizes a way in which the commitment problem enters persuasion. It is neither assumed away, nor does it collapse persuasion to cheap talk. Here, it becomes an endogenous restriction on the information-structure space. Different observability structures correspond to different commitment structures, not by assumption, but through the equilibrium determined by the value function. Instead of justifying commitment or solving the commitment problem, this perspective that connects to friction joins the discussion of how different commitment structures affect the persuasion outcome, which according to Theorem 1, is determined by admissibility. Because the admissibility is endogenous to the value function and microfounded based on the reviewer distribution, this effect is contingent on both the persuasion problem and the observability structure.

Corollary 1. *Enlarging the within-cell tie sets weakly enlarges the purely admissible region. In particular, if* $M_k^\sigma(\mathbf{v}) = \Theta_k$ *for every* k*, then every feasible effective information structure in* $\mathcal{C}_\sigma$ *that induces the selected value vector* $\mathbf{v}$ *is purely admissible.*

Corollary 1 formalizes a simple intuition. Partial observability of the reviewer distribution creates the commitment problem only when hidden substitution of reviewer masses changes the sender's continuation payoff. If all reviewer types inside a cell are tied in contributing to the sender's objective, then deviations within that cell are payoff-irrelevant to the sender, which nullifies the commitment friction. While this is a general corollary, $v_1 = v_2$

in the binary case is the most transparent example. If both signals realize the same payoff, the sender loses the incentive to secretly substitute the reviewer distribution to increase the probability of any of these signals. The same logic applies to the higher dimensional case whenever $(\sigma\mathbf{v})^\top \Pi_\theta p_s$ is constant within a cell.

The corollary also suggests how policy can improve strategy admissibility and thus persuasion effectiveness. Compared to ensuring the observability of certain parts of the reviewer distribution, it may be easier to design contracts and transfers associated with signal realization to neutralize the sender's deviation incentive. To address the commitment problem in persuasion in our indirect context, the idea of mechanism design can be a solution with a slightly different approach, which aims at eliminating incentive incompatibility rather than creating incentive compatibility. If regulation, methodology, or platform design compresses those payoff differences within an observational cell, hidden substitution becomes less consequential. On the other hand, applying Theorem 1 to the two extreme observability structures makes the value function less consequential.

Corollary 2. *If $T\lambda \equiv \lambda$ so that T fully reveals λ, then every direct-persuasion outcome is purely admissible. If the observability structure is completely opaque, so that $T\lambda$ is constant across all $\lambda \in \Delta(\Theta)$, then every purely admissible effective information structure $\Pi \in \mathcal{C}_\sigma$ must belong to the convex hull of the globally maximizing reviewer types,*

$$\operatorname{co}\left\{\Pi_\theta : \theta \in \arg\max_{\theta' \in \Theta} (\sigma\mathbf{v}(\Pi))^\top \Pi_{\theta'} p_s\right\}.$$

In particular, suppose $p_s \in \operatorname{int}(\Delta(\Omega))$. If every information structure that induces nonconstant posterior beliefs also induces a selected value vector with a unique highest component, then every purely admissible outcome is equivalent to cheap talk under complete opacity.

Corollary 2 recovers the conventional benchmarks as the two boundary cases of the observability structure. With the receiver fully observing the reviewer distribution, no hidden

substitution is possible, so the model returns to the standard persuasion benchmark with commitment. With complete opacity, all reviewer types are pooled together. Any informative experiment with a unique highest-value signal gives the sender a profitable hidden deviation to the reviewer type that always sends that signal. Therefore, when informative experiments do not generate top-value ties, the completely opaque boundary leaves only cheap-talk-equivalent outcomes.[8] Most literature focuses only on these boundary cases as information environments, assuming either that commitment is imposed or absent. From a broader perspective, our framework places them on the same continuum. This not only shifts the perspective on commitment from an auxiliary assumption to an equilibrium consequence of observability, but also provides a microfoundation for why markets with hidden reviewer selection, such as ratings shopping, audit opinion shopping, or selective review disclosure, exhibit intermediate commitment structures rather than the all-or-nothing distinction.

Beyond the anchors by Corollaries 1 and 2, the admissibility of the persuasion outcomes returns to the consistency of the value function and the partition structure. In contrast to standard persuasion, the value function here is not only an objective to select the best strategy, but also selects the constraint that makes the strategy feasible. Only when these selections match, as indicated by Theorem 1, is the selected strategy admissible. Using this philosophy, we can perform a point-wise screen to determine the pure-admissibility of each strategy within the whole space, and then optimize the sender's objective on that feasible region.

A point-wise screen to determine admissibility can be a computationally burdensome algorithm, especially when the persuasion problem is high-dimensional. Lemma 2 suggests a finite decomposition of the whole strategy space. Together with Theorem 1, they imply the possibility of reducing the problem of finding the admissible region and the optimal

[8]Informative admissibility can survive under complete opacity only on top-value tie faces, where at least two signals give the sender the same highest value. In that case, the convex hull of the maximizing mappings can contain informative experiments. These cases are the complete-opacity analog of the tie expansion in Corollary 1.

persuasion strategy subject to this endogenous constraint.

Proposition 1. *For a given observability structure* $\{\Theta_k\}_{k=1}^B$*, let* $\mathfrak{G}_k := 2^{\Theta_k} \setminus \{\varnothing\}$ *with* G_k *being an element of* $\mathfrak{G}_k$*, and* $\mathfrak{G} := \prod_{k=1}^B \mathfrak{G}_k$ *with* $G = (G_1, \ldots, G_B)$ *being an element of* $\mathfrak{G}$*. The admissible region can be expressed as*

$$\bigcup_{G \in \mathfrak{G}} \bigcup_{\sigma \in \Sigma} \Bigg(\Big\{ \Pi \in \mathcal{C}_\sigma : (\sigma \mathbf{v}(\Pi))^\top \Pi_\theta p_s \geq (\sigma \mathbf{v}(\Pi))^\top \Pi_{\theta'} p_s, \ \forall \theta \in G_k, \ \forall \theta' \in \Theta_k, \ \forall k \Big\} \bigcap \operatorname{co} \Big\{ \Pi_\theta : \theta \in \bigcup_{k=1}^B G_k \Big\} \Bigg),$$

where $\mathbf{v}(\Pi)$ *denotes the selected sender value vector induced by the receiver's best response to the expected information structure* Π*.*

This decomposition anchors a convex hull, derives the associated admissible region, and sums it over a finite number of possible convex hulls to obtain the entire admissible region. In addition to a more tractable way to characterize the admissible region, Proposition 1 more importantly provides an algorithm to determine the optimal persuasion strategy. This algorithm finds optimal strategies in different finite regimes (G, σ) and compares them to determine the global optimum. In comparison to the algorithm of mapping the purely admissible region and then optimizing within it, this decomposition algorithm is better suited to higher-dimensional environments, where one seeks the optimum, but does not need the full admissible geometry.

The tractability and efficiency of this algorithm is based on finite comparisons among optima in different regimes. Ties between reviewer types enlarge the family of regimes by replacing vertices with maximizing faces, but they do not destroy tractability. In applications, the enumeration can often be cut down further by deleting reviewer types that are never maximizers within the observational cell on a given canonical region. With specific value

functions, the following stylized examples show how partial observability can impact the persuasion outcome when $n = 2$. These binary examples use the perspective of Proposition 1 to explicitly characterize the admissible region and to efficiently identify the constrained optimum.

Examples in the Binary Setting

In the setting with two states ($\Omega = \{\omega_1, \omega_2\}$) and two signals ($\mathcal{S} = \{s_1, s_2\}$) as set up in Section 4, suppose that the receiver observes the proportions of type θ_1 and type θ_2 reviewers separately, but observes only the total mass of the remaining two types of reviewers. Let the selected value vector $\mathbf{v}$ be $(v_1, v_2)^\top$, where v_i denotes the value that the sender obtains upon the realization of the signal that favors the state ω_i, $i = 1, 2$. With these settings, the general theorem then reduces to a one-dimensional comparison inside the observational cell $\{\theta_3, \theta_4\}$. The application of Proposition 1 helps construct a clean admissible region in the strategy space.

Proposition 2. *With binary states $\Omega = \{\omega_1, \omega_2\}$ and the observability structure of $\{\Theta_k\}_{k=1}^3 = \{\{\theta_1\}, \{\theta_2\}, \{\theta_3, \theta_4\}\}$, a persuasion strategy $\big(\pi(s_2|\omega_1), \pi(s_2|\omega_2)\big)$ is purely admissible if and only if either*

$$\pi(s_2|\omega_1) = \pi(s_2|\omega_2),$$

or

$$(v_2 - v_1)\Big[\pi(s_2 \mid \omega_1) + \pi(s_2 \mid \omega_2) - 1\Big] \begin{cases} \geq 0 & \text{if } \pi(s_2 \mid \omega_1) < \pi(s_2 \mid \omega_2) \\ \leq 0 & \text{if } \pi(s_2 \mid \omega_1) > \pi(s_2 \mid \omega_2) \end{cases}, \tag{5}$$

With Proposition 2, the sender's problem is simplified to maximizing the objective $(\sigma\mathbf{v})^\top(\sum_{\theta\in\Theta} \lambda(\theta)\Pi_\theta)p_s$ under the constraint, or geometrically, the wedge condition of (5).

Given this simplified problem, the optimal persuasion strategy depends on the value function characterized by v_i as functions of $(\pi(s_2|\omega_1), \pi(s_2|\omega_2))$, $i = 1, 2$. In the concrete case below, we assume that the prior belief shared by the sender and the receiver is $\Pr(\omega = \omega_1) = 0.4$. With binary states $\{\omega_1, \omega_2\}$, any belief can be fully summarized by the probability assigned to ω_1. We therefore use q to represent this scalar belief in our examples for simplicity. The examples emphasize two different ways in which the same observability friction changes the optimal persuasion strategy. In one case, the sender stays in the same regime and attenuates the information structure; in the other case, the sender switches regimes altogether.

Communication Service (Piecewise Affine Value Function)

Consider a value function

$$V(q) = \begin{cases} 15q - 2 & \text{if } q \in [0, 0.2] \\ -10q + 3 & \text{if } q \in [0.2, 0.3] \\ 5q - 1.5 & \text{if } q \in [0.3, 0.7] \\ -\frac{20}{3}q + \frac{20}{3} & \text{if } q \in [0.7, 1] \end{cases}.$$

This value function may capture a market in which user consumption affects the seller's cost, such as the rental market or the market for communications (internet and phone) service. Users' higher confidence in quality increases demand, but limited capacity makes service provision costly beyond the optimal scale. However, sufficiently strong confidence justifies the user's choice of upgrading to a premium service, which allows the sender to upgrade the capacity to earn higher profit. The resulting sender payoff is therefore non-concave and has two local peaks, as shown in Panel (a) of Figure 1.

Under full observability where $\Theta_k = \{\theta_k\}$ for $k = 1, 2, 3, 4$, the benchmark persuasion strategy is $(\pi(s_2|\omega_1), \pi(s_2|\omega_2)) = (0.3, 0.8)$ and, after signal relabeling, $(\pi(s_2|\omega_1), \pi(s_2|\omega_2)) =$

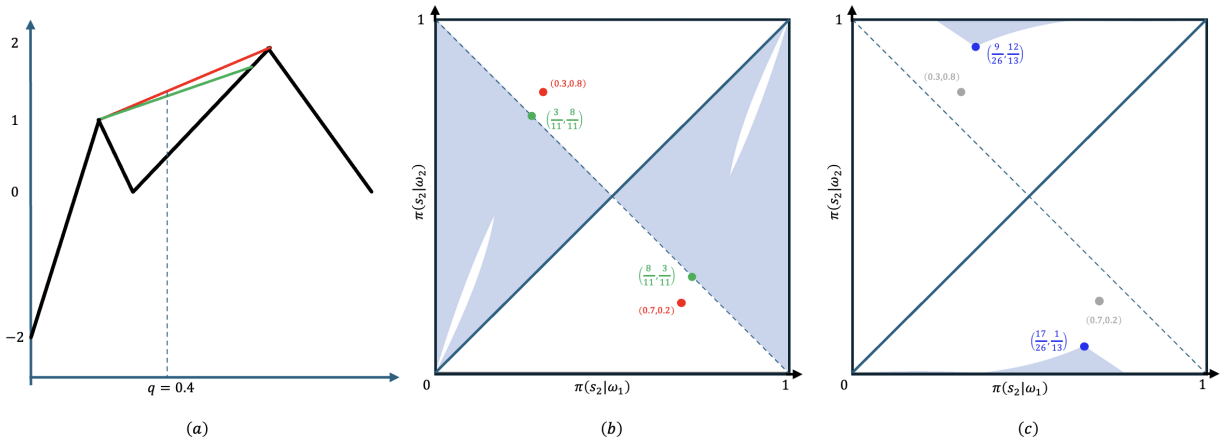

Figure 1: Example 1: Piecewise Affine Value Function

$(0.7, 0.2)$. This strategy potentially generates posterior beliefs $q = 0.2$ and $q = 0.7$, associated with the value $v_2 = 1$ and $v_1 = 2$, respectively. Since $\Pr(\omega = \omega_1) = 0.4$, the sender's expected payoff is $0.6 \times 1 + 0.4 \times 2 = 1.4$. This coincides with the frictionless direct-persuasion benchmark as shown in Panel (a) of Figure 1.

However, under partial observability where the receiver only observes $\lambda(\theta_3) + \lambda(\theta_4)$, the benchmark strategies are not admissible. They induce $v_1 > v_2$ but satisfy $[\pi(s_2|\omega_1) - \pi(s_2|\omega_2)][\pi(s_2|\omega_1) + \pi(s_2|\omega_2) - 1] < 0$. This violates condition (5). Geometrically, the benchmark point lies outside the shaded admissible wedge under $v_1 > v_2$ in Panel (b) of Figure 1.

Under the constraint (5), the sender chooses $q = 0.2$ and $q = 0.64$, which maximizes the persuasion value conditional on $v_1 \geq v_2$, or $q = 0.2$ and $q = 0.85$, which maximizes the persuasion value conditional on $v_1 \leq v_2$. These two options are represented in Panels (b) and (c) in Figure 1. Since the former generates the sender's expected payoff, $\frac{29}{22}$, greater than that generated by the latter, 1, the optimal persuasion strategy under the partial observability constraint is $(\pi(s_2|\omega_1), \pi(s_2|\omega_2)) = (\frac{3}{11}, \frac{8}{11})$ and after signal relabeling, $(\pi(s_2|\omega_1), \pi(s_2|\omega_2)) = (\frac{8}{11}, \frac{3}{11})$.

When constraints are considered, there are two pairs of natural candidates. The first lies within the shaded region in Panel (b) and maximizes the persuasion value conditional on $v_1 \geq v_2$. It is at the boundary of the admissible wedge $(\pi(s_2|\omega_1), \pi(s_2|\omega_2)) = (\frac{3}{11}, \frac{8}{11})$ or, after relabeling, $(\pi(s_2|\omega_1), \pi(s_2|\omega_2)) = (\frac{8}{11}, \frac{3}{11})$. This pair of strategies induces $q = 0.2$ and $q = 0.64$, and the corresponding sender's expected payoff is $\frac{29}{22}$. The second candidate stays within the wedge $[\pi(s_2|\omega_1) - \pi(s_2|\omega_2)][\pi(s_2|\omega_1) + \pi(s_2|\omega_2) - 1] < 0$ and instead maximizes the persuasion value conditional on $v_2 \geq v_1$. This strategy uses $(\pi(s_2|\omega_1), \pi(s_2|\omega_2)) = (\frac{9}{26}, \frac{12}{13})$ or, after relabeling, $(\pi(s_2|\omega_1), \pi(s_2|\omega_2)) = (\frac{17}{26}, \frac{1}{13})$ to induce posterior beliefs $q = 0.2$ and $q = 0.85$. Since both posteriors lie on segments that deliver value 1, the sender's expected payoff is simply 1. Comparing the two admissible candidates, the sender prefers the first.

The interpretation is straightforward. In this example, the sender prefers to preserve the ranking $v_1 \geq v_2$ and move to the nearest admissible region under this condition. The friction does not overturn the direction of persuasion. It forces the sender to make the favorable signal less favorable in order to eliminate the hidden deviation incentive inside the observational cell.

Innovation Path (Step Value Function)

Suppose that the value function takes the form

$$V(q) = \begin{cases} -2 \text{ if } q \in [0, 0.2] \\ 1 \text{ if } q \in [0.2, 0.3] \\ 0 \text{ if } q \in [0.3, 0.7] \\ 2 \text{ if } q \in [0.7, 1] \end{cases},$$

This value function captures the innovation path or a market of products that can be easily pirated. At the initial stage when the project was just introduced, given low market confidence, the project may not be profitable. It earns a moderate return once the

market begins to recognize quality. However, when the confidence of the market continues to accumulate, potential competitors are attracted to the market. Consumers shop for price rather than quality, as they are still not fully convinced about the quality of the product. Therefore, intensified entry drives the rent to 0. Only when the consumer's confidence exceeds the high threshold 0.7 does quality differentiation start to dominate price competition. At this stage, the project becomes highly profitable. In this value function, we keep the overlapping specification at the boundary. At $q \in \{0.2, 0.3, 0.7\}$, the receiver has multiple best responses. In this pure-admissibility example, the fixed sender-favorable pure-selection convention applies at these boundary posteriors.

Under full observability, the benchmark persuasion strategy is again $(\pi(s_2|\omega_1), \pi(s_2|\omega_2)) = (0.3, 0.8)$ and its relabeling $(\pi(s_2|\omega_1), \pi(s_2|\omega_2)) = (0.7, 0.2)$. They induce posterior beliefs $q = 0.2$ and $q = 0.7$, which are associated with $v_2 = 1$ and $v_1 = 2$ and generate the sender's expected payoff 1.4, as shown in Panel (a) of Figure 2. Under partial observability, this benchmark that induces $v_1 > v_2$ remains inadmissible for the same geometric reason as before. They lie within the upper and lower wedges in the square area shown in Panel (b) of Figure 2, which violates (5).

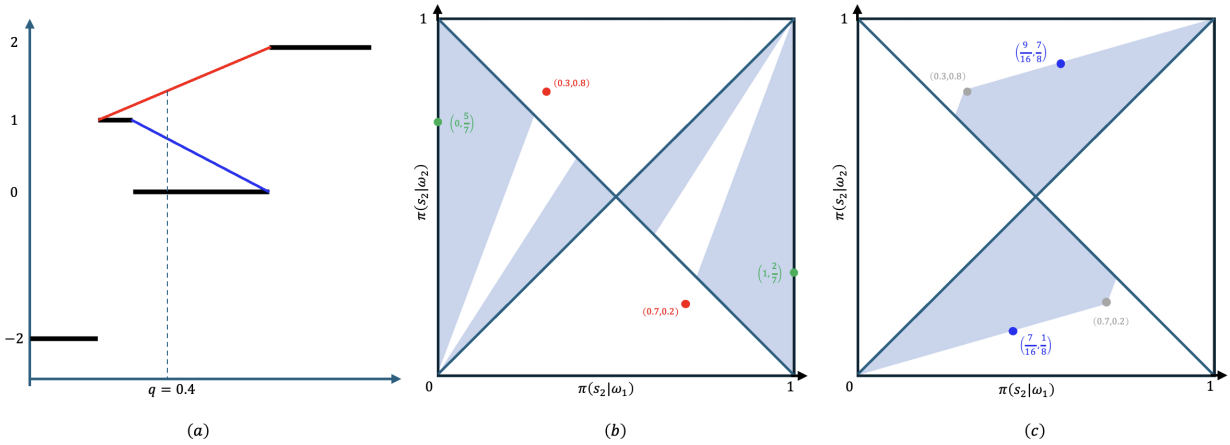

Figure 2: Example 2: Step Value Function

To fix this inadmissibility, the first adjustment keeps the value ranking of $v_1 \geq v_2$ and moves the strategy to the boundary points $(\pi(s_2|\omega_1), \pi(s_2|\omega_2)) = (0, \frac{5}{7})$ and its relabeling $(\pi(s_2|\omega_1), \pi(s_2|\omega_2)) = (1, \frac{2}{7})$. These strategies induce posteriors $q = 0$ and $q = 0.7$ and yield the sender's expected payoff $\frac{2}{7}$. The second admissible adjustment instead switches to the optimization conditional on $v_2 \geq v_1$ and uses $(\pi(s_2|\omega_1), \pi(s_2|\omega_2)) = (\frac{9}{16}, \frac{7}{8})$ and its relabeling strategy $(\pi(s_2|\omega_1), \pi(s_2|\omega_2)) = (\frac{7}{16}, \frac{1}{8})$. This solution generates an expected payoff $0.75 > \frac{2}{7}$ for the sender and, therefore, dominates the first. In this example, the cheapest way to restore admissibility is not to stay in the original regime and attenuate the benchmark strategy. It is to switch to the regime that requires $v_2 \geq v_1$. With this change, the original bad signal becomes a good one, since the other signal generates an even lower value, which measures the price of observability becoming incomplete.

The two examples compare the two margins by which partial observability changes the optimal persuasion strategy. Here, the inadmissibility of the benchmark strategy is characterized by the mismatch between the characteristics of the value vector and the strategy location. This mismatch in the binary case is captured by condition (5), which naturally points towards two solutions to the inadmissibility. The sender can remain in the same regime and move to the nearest admissible region as in the first example. If this is too costly, as in the second example, she changes the regime itself. In both cases, partial observability does not eliminate persuasion. It removes those information structures whose implementation would require the sender to retain hidden reviewer types that are dominated inside an observational cell.

6 Mixed Strategy and Expected Admissibility

Pure admissibility represents the ex post implementability of a strategy on a fixed fiber. Therefore, once the sender chooses the reviewer distribution that induces a relevant static continuation game, we take a pure-strategy equilibrium in this continuation game as the

benchmark notion of admissibility and assume that the game prioritizes this implementation once it exists. However, a fixed observable statistic $z = T\lambda$ can also induce a *de facto* static continuation game on $\Lambda(z)$ with no pure-strategy equilibrium. In this case, the continuation game is not silent. Under the usual existence conditions, a mixed-strategy equilibrium exists, and the corresponding information structures define a residual notion of admissibility, which is revealed only when pure admissibility fails.

We define the concept of **expectedly admissible** based on this fact. In a pure-strategy equilibrium, a value vector generated by a receiver action rule selects the reviewer types that are incentive-compatible within each observational cell. In a mixed-strategy equilibrium, the receiver mixes over action rules, and the sender responds to the resulting expected value vector.

Fix a fiber $\mathcal{P}(z)$. An action rule is a mapping $\alpha : \mathcal{S} \to \mathcal{A}$, or equivalently $\alpha \in \mathcal{A}^{\mathcal{S}}$. Let $U(\Pi, \alpha)$ denote the receiver's ex ante payoff from action rule α when the receiver expects the information structure Π. We keep U as a separate object because alternative specifications of receiver behavior may change this ex ante payoff while leaving the state-action payoff u unchanged.[9] On this basis, define the receiver's best response correspondence to the expected information structure Π:

$$BR_r(\Pi) := \arg\max_{\alpha \in \mathcal{A}^{\mathcal{S}}} U(\Pi, \alpha).$$

Unlike pure admissibility, which uses the fixed pure-selection convention to obtain $\mathbf{v}(\Pi)$, expected admissibility allows the receiver to randomize over action rules in $BR_r(\Pi)$.

Theorem 2 (Expected Admissibility)**.** *Fix T and a feasible observable statistic z with nonempty fiber $\mathcal{P}(z)$. Suppose that the receiver's action set $\mathcal{A}$ is finite and that, for each action rule $\alpha \in \mathcal{A}^{\mathcal{S}}$, the receiver's ex ante payoff $U(\Pi, \alpha)$ is affine in Π. If the fiber contin-*

[9] In the Bayesian benchmark without additional behavioral terms, $U(\Pi, \alpha) = \sum_{\omega \in \Omega} \sum_{s \in \mathcal{S}} p_r(\omega)\pi(s|\omega)u(\omega, \alpha(s))$, which is affine in Π for each fixed α.

uation game induced by z has no pure-strategy equilibrium, then there exist an information structure $\Pi^E \in \mathcal{P}(z)$ and a nondegenerate receiver mixture $\tau^* \in \Delta(\mathcal{A}^{\mathcal{S}})$, $|\operatorname{supp}\tau^*| \geq 2$, such that $\operatorname{supp}\tau^* \subseteq BR_r(\Pi^E)$ and

$$\Pi^E \in \sum_{k=1}^{B} z_k \operatorname{co}\left\{\Pi_\theta : \theta \in \arg\max_{\theta\in\Theta_k} (\sigma\bar{\mathbf{v}}(\tau^*))^\top \Pi_\theta p_s\right\},$$

where $\mathbf{v}(\alpha)$ is the sender's value vector induced by α, $\bar{\mathbf{v}}(\tau^*) := \sum_{\alpha\in\mathcal{A}^{\mathcal{S}}} \tau^*(\alpha)\mathbf{v}(\alpha)$, and σ is the canonical region that contains Π^E. Any such Π^E is defined as ***expectedly admissible on the fiber*** *$\mathcal{P}(z)$.*

Compared to pure admissibility, the receiver's behavior plays a more direct role in expected admissibility. For pure-admissibility results, the exact cardinal specification of U matters only through the receiver's best-response correspondence. For expected admissibility, it also matters through the receiver's indifference set for given information structures, because according to Theorem 2, this is where the receiver can mix. To locate where this indifference appears, additional specification of U is needed, especially how it varies between different information structures outside of the best-response region.

Assuming that there is no additional behavioral punishment for a mismatch between the information structure and the action rule, the affine specification of $U(\Pi, \alpha)$ is a convenient benchmark. An important implication of this property is that the receiver evaluates a sender mixture through its barycenter in Π-space, so any sender mixture has an equivalent pure strategy on the fiber. Thus, the central object in expected admissibility is the receiver's mixture. In our model, any mixed-strategy equilibrium can be sustained by the sender's pure strategy. This observation not only clarifies the mechanism that sustains the equilibrium, but also simplifies both the location of the expectedly admissible region and the determination of the associated persuasion value.

The comparison between Theorems 1 and 2 reveals how mixed strategies can restore

admissibility when pure admissibility fails and no pure-strategy equilibrium on the fiber conceals the mixed-strategy equilibrium. Admissibility requires a match between the strategy Π and the endogenous convex hull in the strategy space. The receiver's mixture over action rules creates an expected value vector that admits additional mappings that were not selected by any pure receiver action rule. Geometrically, it expands the convex hull to improve matching. This is the mixed analog of the tie-expansion mechanism of Corollary 1. There, if a value vector ties several reviewer types in a cell, the selected convex hull expands to facilitate pure admissibility. Here, the same expansion is generated endogenously by the receiver's mixture. The difference is that the resulting admissibility is only expected, relying on the receiver's indifference and randomization.

This logic also explains why expected admissibility is rare. Expected admissibility requires receiver indifference that is relevant for the sender's payoff. With a finite receiver action set, such indifference appears only on boundaries of the receiver's best-response regions. In the reduced-form value correspondence, these are points where different sender values may be available for the same posterior, potentially implying discontinuity in the value function. If the value function is continuous along the relevant fiber, this indifference usually does not reveal expected admissibility. Particularly, when endpoint incentives point in opposite directions, continuity tends to generate an intermediate information structure with a level value segment, which sustains a pure-strategy equilibrium and conceals the mixed one. If the receiver's action-rule space is convexified so that the receiver's mixed strategy itself becomes a feasible pure strategy, the same mixed object is again absorbed into pure admissibility. Therefore, expected admissibility is relevant only in the residual case where no pure-strategy equilibrium exists and the receiver still has a payoff-relevant indifference set on the fiber.

Adding to the rarity of expected admissibility is the robustness of cheap talk. As long as a fiber contains an uninformative information structure, cheap talk becomes a natural pure-strategy equilibrium candidate because all signals induce the same posterior and the same

sender value. Under our equilibrium-selection convention, this pure-strategy equilibrium conceals any mixed-strategy equilibrium on the same fiber. This gives a different perspective on cheap talk. Cheap talk is dangerous not because persuasion must always collapse to it, but because it is often the most robust admissible point when commitment is weakened by partial observability. In the binary example where the receiver observes only the proportions of type θ_3 and type θ_4 reviewers, the robust admissible region is the diagonal of the square space. Since all fibers cross this diagonal, expected admissibility does not survive. This perspective also suggests a policy margin. If improving observability directly is difficult or costly, a policy designer may instead try to change the sender's payoff through transfers and institutional rules so that more informative strategies become pure-strategy equilibria and no longer lose to the uninformative ones within the fiber.

Theorem 2 establishes existence under the maintained no-pure-equilibrium premise. It also suggests an algorithm for locating expectedly admissible regions. The algorithm first screens fibers, removes those with pure-strategy equilibria, and then examines receiver-indifference sets on the remaining fibers. However, the theorem itself does not provide a characterization of which indifference sets can support a mixed-strategy equilibrium on this fiber. Theorem 3 rewrites the continuation problem as a score-balancing condition over the extreme points of the fiber. This gives a sharper principle for finding the relevant equilibrium information structures.

Theorem 3. *Suppose that the receiver's action set $\mathcal{A}$ is finite. Fix T, a nonempty fiber $\mathcal{P}(z) := \{\Pi(\lambda) : \lambda \in \Lambda(z)\}$ and a candidate information structure $\Pi \in \mathcal{P}(z)$. Let $\mathrm{Ext}(\mathcal{P}(z)) = \{\widetilde{\Pi}_1, \ldots, \widetilde{\Pi}_K\}$ denote the extreme points of the fiber image. For each $\alpha \in BR_r(\Pi)$, define the score vector $\mu(\alpha) := \left((\sigma\mathbf{v}(\alpha))^\top \widetilde{\Pi}_1 p_s, \ldots, (\sigma\mathbf{v}(\alpha))^\top \widetilde{\Pi}_K p_s\right) \in \mathbb{R}^K$, where σ is the canonical region containing Π. Then Π can be sustained as an equilibrium outcome of the* de facto *static continuation game associated with $\mathcal{P}(z)$ if and only if there exists a nonempty set*

$\Phi \subseteq \{1, \ldots, K\}$ *such that*

$$\Pi \in \text{co}\{\widetilde{\Pi}_j : j \in \Phi\} \tag{IN}$$

and

$$\min_{\alpha \in BR_r(\Pi)} \beta^\top \mu(\alpha) \leq 0 \qquad \forall \beta \in \mathbb{R}^K \text{ such that } \beta^\top \mathbf{1} = 0 \text{ and } \beta_\ell \geq 0 \ \forall \ell \notin \Phi. \tag{SB}$$

If the continuation game based on $\mathcal{P}(z)$ has no pure-strategy equilibrium, then any Π satisfying (IN) and (SB) is expectedly admissible on the fiber.

Theorem 3 gives a general condition under which an information structure induces the value vector that qualifies its admissibility. Condition (IN) selects a face of the fiber that contains Π. Condition (SB) asks whether the receiver's best-response action rules can be mixed so that the extremes in this selected face receive the same highest score, while all excluded extremes receive weakly lower scores. Given Condition (IN), pure admissibility is the special case in which a single action rule $\alpha \in BR_r(\Pi)$ already satisfies Condition (SB).

If admissibility is not possible, there is a tension between conditions (IN) and (SB) built into the value function. The previous discussion indicates that additional receiver strategies α ease this tension. Theorem 3 goes further to detail this mechanism. For a target information structure, the value vector induced by the receiver's best responses may assign too low a score to the extremes in Φ relative to complementary extremes, creating profitable deviations for the sender captured by some $\beta \in \mathbb{R}^K$. Another receiver action rule may have a different weakness. However, these different action rules cover each other's weaknesses. As a result, the receiver's mixture can raise the relative score of the selected extremes in Φ and suppress that of the remaining ones. In another case, the mixture may expand Φ. These two mechanisms, where the former relaxes condition (SB) for given condition (IN) and the latter goes the other way around, both relax the tension between the two conditions to foster expected admissibility.

The mechanism is especially straightforward in the binary case, where extremes in the fiber cannot induce the receiver's indifferent best responses in the absence of pure admissibility. A fiber has two endpoints. If there is no pure admissibility at either endpoint, the receiver's best-response rule associated with each endpoint makes the sender prefer the opposite endpoint. If both action rules are receiver best responses to some interior information structure, this crossing of endpoint rankings implies condition (SB). The receiver can mix them to equalize the two endpoint scores. Once the endpoints are equalized, every information structure on that fiber gives the sender the same expected payoff. Since reweighting the probabilities of signals does not change her expected payoff, the sender will choose the information structure that makes both action rules the receiver's best responses, which sustains the mixed-strategy equilibrium.

Theorems 2 and 3 also clarify the role of the receiver's payoff function. The receiver's payoff locates the indifference sets where the mixture mechanism can work. In this sense, expected admissibility requires more information about receiver behavior than pure admissibility. Different payoff specifications may change which information structure is supported, or whether the benchmark becomes purely admissible. However, once an indifference set and a supporting mixture are fixed, the effect of partial observability on persuasion effectiveness is determined by the induced value vectors and the observability structure, summarized by the score vectors $\mu(\alpha)$. Therefore, in this conditional sense, the effect captured in the following corollary does not rely on a particular cardinal specification of U and is robust across receiver-payoff specifications.

Corollary 3. *Let* $\Pi^* \in \arg\max_{\Pi \in \mathcal{P}} \left(\sigma \mathbf{v}(\Pi)\right)^\top \Pi p_s$ *be a benchmark persuasion strategy located in canonical region* σ*, where* $\mathbf{v}(\Pi^*)$ *is the selected sender-favorable value vector at* Π^**. Suppose that* $\Pi^* p_s$ *has full support. If* Π^* *is not purely admissible but is expectedly admissible on the fiber* z*, supported by receiver mixture* τ^**, then*

$$(\sigma \bar{\mathbf{v}}(\tau^*))^\top \Pi^* p_s < (\sigma \mathbf{v}(\Pi^*))^\top \Pi^* p_s.$$

Corollary 3 highlights an important implication that expected admissibility can rescue implementability but comes with a discount on the payoff associated with the information structure. If the optimal benchmark strategy is not purely admissible, the only possible way to sustain the same benchmark information structure on its fiber is through the receiver's action-rule mixture. This mixture restores admissibility by suppressing the high value that made the hidden deviation attractive to the sender. Therefore, expected admissibility does not restore the full-commitment payoff if pure admissibility cannot. This discount represents a different channel through which partial observability affects persuasion. Pure inadmissibility can remove information structures from the feasible set. Expected admissibility can bring some back, but only under an expected value vector generated by receiver mixing. The friction does not necessarily render the optimal benchmark information structure inadmissible, but when the benchmark is not purely admissible and has full signal support, the information cost is unavoidable.

Example

The binary examples from the preceding section illustrate how restrictive expected admissibility is. In the piecewise affine example, the continuation games always admit pure-strategy equilibria,[10] so expected admissibility, as a residual concept, is not revealed. With the step value function, expected admissibility appears only for two observability structures: the receiver observes one constant-signal reviewer type (θ_3 or θ_4) and one state-contingent reviewer type (θ_1 or θ_2), while the remaining two reviewer types are pooled. Here, we focus on the case where the receiver observes the proportions of type θ_2 and θ_4 reviewers, while

[10]Since this specific value function is continuous, as shown in Panel (a) of Figure 1, if no pure-strategy equilibrium exists at the endpoints, the two endpoint information structures must carry opposite hidden-deviation incentives. Then, by continuity, there exists an information structure in between that induces a level value segment and neutralizes the sender's deviation incentive, sustaining a pure-strategy equilibrium.

type θ_1 and θ_3 reviewers are pooled. The symmetric case is obtained by relabeling signals and states.

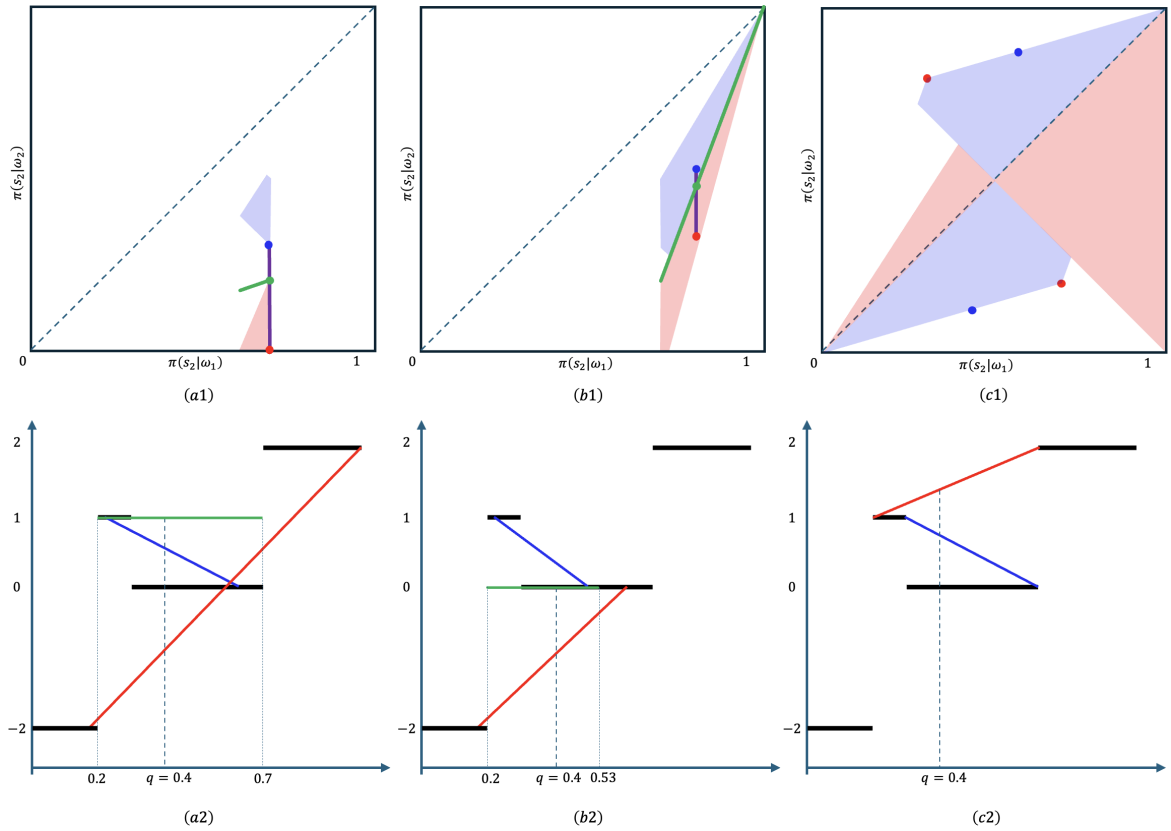

Figure 3: Example: Expected Admissibility and Pure Admissibility

In Figure 3, Panels (a1) and (b1) show the extremes of the fibers on which there is no pure-strategy equilibrium. As the representative fibers in both panels illustrate, these fibers are vertical line segments, with lower and upper endpoints lying in the red and blue regions, respectively. According to Theorems 2 and 3, expected admissibility requires the receiver's indifference conditional on the sender's equilibrium strategy. In this example, the receiver's indifference occurs only when at least one induced posterior belief is at a discontinuity of the value function $\{0.2, 0.3, 0.7\}$, as shown in Panels (a2) and (b2). These expectedly admissible information structures, which are the sender's equilibrium strategies,

are represented by green line segments in Panels (a1) and (b1).

All expectedly admissible information structures on the branch in Panel (a1) generate a payoff of 1 for the sender. Panel (a2) gives a representative fiber on this high-value branch. This representative fiber has endpoints

$$\big(\pi(s_2|\omega_1), \pi(s_2|\omega_2)\big) = (0.7, 0) \quad \text{and} \quad \big(\pi(s_2|\omega_1), \pi(s_2|\omega_2)\big) = (0.7, 0.3).$$

They induce the posterior pairs $(q_1, q_2) = (\frac{1}{6}, 1)$ and $(q_1, q_2) = (\frac{2}{9}, \frac{14}{23})$, respectively, as represented by the two line segments in Panel (a2). At either endpoint, the sender has an incentive to move secretly to the other endpoint, because this increases the probability of the signal associated with the higher value. Therefore, no pure-strategy equilibrium exists on this fiber.

However, the fiber contains the information structure

$$\Pi^E = \big(\pi(s_2 \mid \omega_1), \pi(s_2 \mid \omega_2)\big) = (0.7, 0.2),$$

which induces $(q_1, q_2) = (0.2, 0.7)$. Both boundary posteriors are discontinuities of the value function, where the receiver's indifference appears. Under the sender-favorable convention, the receiver can choose a mixed action rule that gives the sender's expected value 1 after both signals. Specifically, at posterior $q = 0.7$, the receiver mixes between the actions yielding values 0 and 2 with equal probability, while at posterior $q = 0.2$, he selects the action yielding value 1. Because the expected value is the same after both signal realizations, changing the probability of the signals along the fiber no longer changes the sender's payoff. The hidden deviation incentive is neutralized.

In contrast, all expectedly admissible information structures on the branch in Panel (b1) yield payoff 0 for the sender. Panel (b2) gives a representative fiber on this low-value branch.

The two endpoints in this fiber are

$$\big(\pi(s_2|\omega_1), \pi(s_2|\omega_2)\big) = \left(0.8, \frac{1}{3}\right) \quad \text{and} \quad \big(\pi(s_2|\omega_1), \pi(s_2|\omega_2)\big) = \left(0.8, \frac{8}{15}\right),$$

inducing the posterior pairs $(q_1, q_2) = (\frac{1}{6}, \frac{8}{13})$ and $(q_1, q_2) = (\frac{2}{9}, 0.5)$, respectively, as shown in Panel (b2). As in the high-value branch, each endpoint gives the sender an incentive to deviate secretly to the other endpoint, so pure admissibility also fails on the fiber.

The fiber contains the information structure

$$\Pi^E = \big(\pi(s_2|\omega_1), \pi(s_2|\omega_2)\big) = \left(\frac{4}{5}, \frac{7}{15}\right),$$

which induces $(q_1, q_2) = (0.2, \frac{8}{15})$, shown in Panel (b2). The first posterior is at the discontinuity $q = 0.2$, while the second posterior lies in the interior of the zero-value region. Hence, given the receiver's best response after signal s_2, the sender's value is fixed at 0. At posterior $q = 0.2$, the receiver mixes the actions resulting in -2 and 1 for the sender's payoff, choosing the latter with probability 2/3. The sender's expected payoff after signal s_1 is then also 0, matching the value after signal s_2. This again neutralizes her incentive to secretly change the signal probabilities along the fiber.

As a reference, Panels (c1) and (c2) return to pure admissibility. According to the preceding section, a strategy must lie in the region matching its color to be purely admissible. This principle implies that the benchmark strategies remain outside the designated admissible region, so the full-commitment benchmark outcome cannot be implemented in pure strategies. In this case, the best purely admissible strategies remain

$$\big(\pi(s_2|\omega_1), \pi(s_2|\omega_2)\big) = \left(\frac{9}{16}, \frac{7}{8}\right) \quad \text{and} \quad \big(\pi(s_2|\omega_1), \pi(s_2|\omega_2)\big) = \left(\frac{7}{16}, \frac{1}{8}\right),$$

which induce posterior pairs $(q_1, q_2) = (0.7, 0.3)$ and $(q_1, q_2) = (0.3, 0.7)$, up to signal relabeling. These strategies generate the sender's expected payoff 0.75. In comparison, the

best expectedly admissible branch generates payoff 1, while the full-commitment benchmark generates $\frac{7}{5}$. In this example, the sender's payoff ranking is therefore $\frac{7}{5} > 1 > \frac{3}{4}$.

This ranking summarizes the role of expected admissibility. It can restore the admissibility of an information structure when the pure-strategy equilibrium fails on the fiber in which it is located, and it may even outperform the best purely admissible strategy. However, it does not restore the full-commitment benchmark. The receiver's mixture saves implementability by neutralizing the sender's hidden-deviation incentive, but the same mixture changes the sender's effective value vector. Thus, expected admissibility is a remedy for the commitment problem, not a complete solution to it.

The remedy is also fragile in a way that pure admissibility is not. Expected admissibility relies on the receiver's indifference and on a precise randomization device. The payoff it delivers is an expected payoff under the receiver's mixture, not a payoff secured signal-by-signal under a pure action rule. In our model, if the receiver's mixture is properly implemented, the sender compares strategies based on expected payoffs. In practice, however, if there are issues with the receiver's randomization device, the sender may still prefer a purely admissible strategy despite its lower expected value. This is why expected admissibility remains a residual concept in our framework. If a policy maker wants to preserve the expected admissibility of an information structure, it may be necessary to provide a reliable randomization or tie-breaking device that makes the neutralization credible.

7 Robust Observability Structure

Given a specific persuasion problem, the observability structure determines how reviewer types are pooled and, therefore, which effective information structures are admissible. As a feasible constraint on persuasion outcomes, the observability structure is more stable than persuasion problems. It can matter across many persuasion problems with different primitives, which in our environment are the sender's value vector $\mathbf{v}$, the policy target

strategy $\Pi^\dagger$,[11] and the sender's prior p_s. A natural question is therefore, which observability structures can keep the target persuasion outcome admissible as these primitives vary.

This robustness can be approached from two perspectives. The first is an *ex ante* design problem: the observability structure is chosen or refined before the persuasion profile is known, but the primitives are revealed by the time persuasion takes place. The second is an *on-path* robustness problem: even at the persuasion stage, some primitive remains private, particularly the sender's prior belief. The first question asks which observability structure is needed to preserve flexibility across future persuasion problems; the second asks whether asymmetric information about certain primitives matters under a fixed observability structure.

The first perspective is close to the actual institutional design. Review platforms and policy makers typically choose a durable disclosure rule before any particular market activities or persuasion problem are known, and then apply that rule across many heterogeneous environments. For example, in the buyer-seller scenario, the review system is established to adapt to the customer's varying willingness to pay, which not only determines the belief cutoff that motivates customers to purchase (benchmark Π^* as the policy target $\Pi^\dagger$), but also the sender's profit from the induced purchase action ($\mathbf{v}$). In hotel reviews, Expedia historically restricted reviews to customers who booked through the platform, while TripAdvisor allowed anyone to post. These platform-design distinctions remain persistent despite the design leading to consistent review manipulation patterns (Mayzlin et al. 2014).

According to Corollary 2, robustness is trivial under full observability. If the receiver fully observes the reviewer distribution or, technically, reviewer types are partitioned into singletons, the admissible region coincides with the full set of effective information structures regardless of primitive values. However, this benchmark may seem too strict. The interesting question then is how much of that benchmark can be retained when the designer cannot, or

[11] The target $\Pi^\dagger$ may coincide with the full-commitment benchmark Π^*, but it may also be a chosen substitute when the benchmark is too costly to preserve.

does not want to, refine the observability structure all the way to singletons.

Theorem 4. *Fix $\hat{i} \in \{1, 2, \ldots, n\}$ and let $\Theta^{\hat{i}} \subset \Theta$ be the subset of deterministic reviewer types who choose $s_{\hat{i}}$ for at least one state, so that $|\Theta^{\hat{i}}| = n^n - (n-1)^n$. Define the polytope*

$$\mathcal{P}^{\hat{i}} \equiv \Big\{ \Pi \in \mathcal{P} : \sum_{j=1}^{n} \pi(s_{\hat{i}} \mid \omega_j) \geq 1 \Big\}.$$

Then $\mathcal{P}^{\hat{i}}$ is a polytope whose vertices are the deterministic response matrices induced by $\Theta^{\hat{i}}$. Furthermore, there exists a proper fundamental region $\mathcal{D} \subset \mathcal{P}^{\hat{i}}$, and every $\Pi \in \mathcal{D}$ can be implemented by some $\lambda \in \Delta(\Theta^{\hat{i}})$ via

$$\pi(s \mid \omega) = \sum_{\theta \in \Theta^{\hat{i}}} \lambda(\theta) \mathbf{1}\{s = \gamma(\theta, \omega)\}.$$

Theorem 4 shows that the observability of all n^n reviewer types is sufficient, but not necessary, to recover full admissibility up to permutation. In addition, this theorem further reduces this sufficiency to $n^n - (n-1)^n$ types that use one designated signal at least once. The omitted $(n-1)^n$ types are precisely those that never use that signal, which are redundant as shown in the constructive proof. After a suitable permutation of signal labels, every effective information structure can be represented using these $n^n - (n-1)^n$ reviewer types that employ the designated signal somewhere.

By this theorem, the designer does not need to certify every possible reviewer mapping. What matters is to separate a strategically chosen spanning family. Put differently, a robust design does not need to reveal the whole type space. It only needs enough singleton cells to canvas a proper fundamental region. Compared to all n^n reviewer types, this reduction to $n^n - (n-1)^n$ is substantial. The fraction of deterministic reviewer types that may remain

pooled is

$$\left(1-\frac{1}{n}\right)^n.$$

Numerically, this corresponds to a reduction that rises from 25% when $n = 2$ to about $e^{-1} \approx 36.8\%$ as $n \to \infty$. Thus, the reduction is modest in the binary case but quantitatively significant once n is larger.

The theorem is especially transparent in the binary case. When $n = 2$, even without θ_4 (the picky) reviewers who always send s_2, the sender can still implement any experiment up to relabeling by letting s_1 play the role of the "good" signal ($\pi(s_2|\omega_2) \geq \pi(s_2|\omega_1)$) when she needs to maximize the likelihood of the good signal, and play the role of the "bad" signal ($\pi(s_2|\omega_2) \leq \pi(s_2|\omega_1)$) when she needs to maximize the probability of the bad signal. This exactly matches both the minimal vertex requirement (Carathéodory) $n(n-1)+1 = 3$ and the bound $n^n - (n-1)^n = 3$.[12] In that sense, the construction is sharp in the binary environment. But precisely because the binary geometry is tight, observing three out of four reviewer types is equivalent to full observability, as the total mass of all reviewers is normalized to 1. Nevertheless, additional information about the persuasion problem allows for further relaxation of the observability structure.

Theorem 4 is primitive-independent and therefore conservative. Given particular space structures, such as those that come with the binary case, once the designer knows even one piece of the persuasion environment, the observability requirement can be relaxed further.

Proposition 3. *In the binary case where $n = 2$, fix a target persuasion outcome and let $\Pi^\dagger$ be a labeled representative of it. If the policy designer knows either $\Pi^\dagger$ or the relevant ordering between v_1 and v_2, then there exists a three-cell observability structure, partitioning the four reviewer types into three cells, such that a relabeling of $\Pi^\dagger$ is purely admissible. Hence the target persuasion outcome is sustained as an equilibrium.*

[12] Essentially, Theorem 4 is a constructive upper bound. Since $\mathcal{P}$ has dimension $(n-1)n$, any (full-dimensional) robust design that covers interior points needs at least $(n-1)n+1$ singleton types by Carathéodory Theorem. These two bounds coincide when $n = 2$, but not in general.

Proposition 3 not only strengthens the implication that a partially observable reviewer distribution is robust to varying primitives in the persuasion problem regarding admissibility. It also implies that, in the binary case, even Theorem 4 is not necessary for this robustness once the sender has some information about the persuasion problem. In particular, if either the target experiment $\Pi^\dagger$ or the relevant value vector $\mathbf{v}$ is already known, one nontrivial pool can be tolerated without losing the target outcome.

In our innovation path example, partial observability changes the optimal persuasion strategy because the observability structure renders the benchmark optimal persuasion strategy inadmissible. However, when we adopt another observability structure, if the receiver either fully observes the proportions of reviewers of type θ_1 and type θ_4, or if he observes the proportions of reviewers of type θ_2 and type θ_3 reviewers, the benchmark becomes admissible up to permutations.

Figure 4 shows the geometry. Panels (a) and (b) shade the admissible regions for these two observability structures in Π-space, which include at least one of the optimal persuasion strategies. More importantly, the graphs also illustrate two distinct mechanisms behind Proposition 3. Panels (a) and (b) correspond to the first route, in which the designer knows the target experiment $\Pi^\dagger$ to admit. The darker region where Π satisfies $[\pi(s_2|\omega_1)+\pi(s_2|\omega_2)-1][\pi(s_2|\omega_2)-\pi(s_2|\omega_1)] \geq 0$ is the intersection of two possible convex hulls. Hence, any Π in these darker regions is admissible regardless of which pooled reviewer the sender prefers. In this case, the remaining information about value vector $\mathbf{v}$ and prior p_s becomes irrelevant for admissibility.

Panels (c) and (d) show the second mechanism, in which the designer does not know $\Pi^\dagger$, but does know the value ranking $v_2 \leq v_1$. Under this restriction, the darker regions in Panels (c) and (d) exhaust the full binary experiment space up to permutation. Therefore, once the designer knows $v_2 \leq v_1$, the information about $\Pi^\dagger$ is no longer needed, as it must lie in the darker region of one of the two panels up to relabeling.

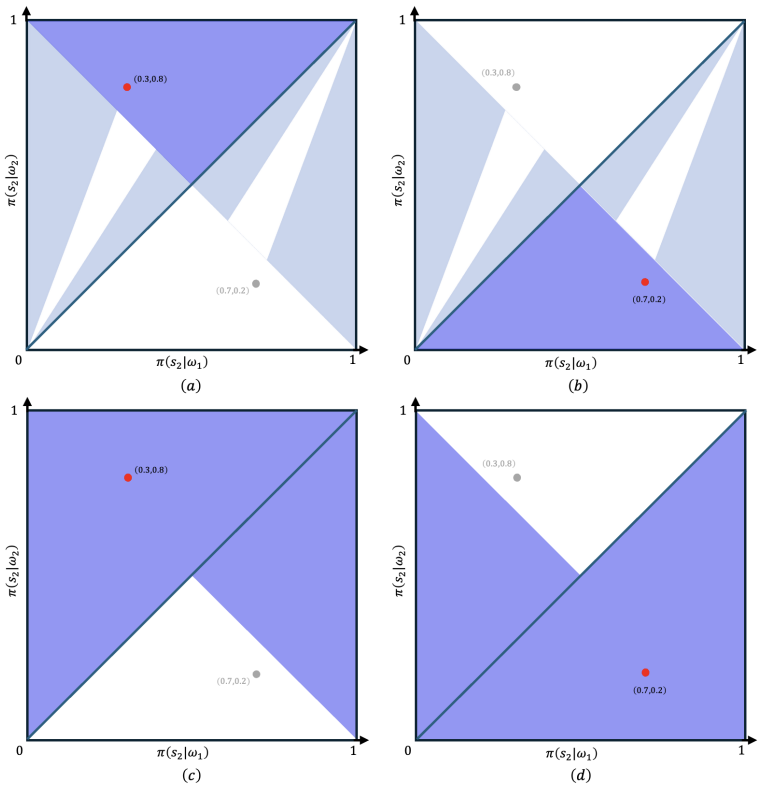

Figure 4: A Robust Design of Observability Structure

Figure 4 also emphasizes the necessity of knowing some correct information. If the designer tries to use the first mechanism with the wrong target experiment, so that the true $\Pi^\dagger$ satisfies the opposite condition $\left[\pi(s_2|\omega_1) + \pi(s_2|\omega_2) - 1\right]\left[\pi(s_2|\omega_2) - \pi(s_2|\omega_1)\right] < 0$, then this $\Pi^\dagger$ may fall within the blank region in Panels (a) and (b), and the chosen observability structure will fail to sustain the target outcome. Likewise, if the designer tries to use the second mechanism with the wrong value ranking, so that in fact $v_2 > v_1$, then the admissible region in Panels (c) and (d) collapses to the diagonal, and only cheap talk remains. Therefore, in the binary case, the designer cannot identify a robust three-cell observability structure without at least one correct piece of information. If both $\Pi^\dagger$ and the ranking of v_i are unknown, the designer may choose an observability structure under which the target persuasion outcome is not admissible.

In addition to the structure design stage, some primitives can remain private even during persuasion, which creates a different question of robustness. In a direct persuasion problem with the full-commitment assumption, the sender's prior belief p_s is irrelevant information to the receiver. However, once the commitment assumption is imperfect, p_s predicts which latent deviation the sender would prefer. When several reviewer types are pooled by the observability structure, the sender's private p_s translates into the receiver's uncertainty about which reviewer within the pool dominates in the sender's preference. In that sense, the sender's prior becomes essential for the receiver to infer the actual information structure in persuasion and for determining which persuasion outcomes can be sustained as equilibrium. A common way to avoid this problem is simply to assume that p_s is publicly known. This assumption is widely applied in the relevant literature, although it may not be universally justified in real world practice. However, the next theorem shows that this is stronger than necessary. Under a restricted but economically meaningful class of observability structures, the receiver's ignorance of p_s does not jeopardize the admissible set of information structures as a constraint that determines the equilibrium.

Theorem 5. *A subfamily $\widetilde{\Theta} \subseteq \Theta$ is universally p_s-robust, namely, for every fixed value vector $\mathbf{v} \in \mathbb{R}^n$ and every pair $\theta', \theta'' \in \widetilde{\Theta}$, $\mathbf{v}^\top (\Pi_{\theta'} - \Pi_{\theta''})\, p_s$ does not change sign as p_s varies over $\Delta(\Omega)$ if and only if there exist a subset $\widetilde{\Omega} \subseteq \Omega$ and a background mapping $\bar{\gamma} : \Omega \setminus \widetilde{\Omega} \to \mathcal{S}$ such that all types in $\widetilde{\Theta}$ agree with $\bar{\gamma}$ on $\Omega \setminus \widetilde{\Omega}$, and one of the following two cases holds:*

(a) There exist two distinct signals $s^+, s^- \in \mathcal{S}$ and, for each $\theta \in \widetilde{\Theta}$, a set $\Omega_\theta \subseteq \widetilde{\Omega}$ such that

$$\gamma(\theta, \omega) = \begin{cases} \bar{\gamma}(\omega), & \omega \notin \widetilde{\Omega}, \\ s^+, & \omega \in \Omega_\theta, \\ s^-, & \omega \in \widetilde{\Omega} \setminus \Omega_\theta, \end{cases}$$

and the family $\{\Omega_\theta : \theta \in \widetilde{\Theta}\}$ satisfies $\Omega_{\theta'} \subseteq \Omega_{\theta''}$ or $\Omega_{\theta''} \subseteq \Omega_{\theta'}$ for any $\theta' \neq \theta''$.

(b) For every $\theta \in \widetilde{\Theta}$, there exists a signal $s_\theta \in \mathcal{S}$ such that

$$\gamma(\theta, \omega) = \begin{cases} \bar{\gamma}(\omega), & \omega \notin \widetilde{\Omega}, \\ s_\theta, & \omega \in \widetilde{\Omega}. \end{cases}$$

A direct implication of Theorem 5 is that if every observational cell pools only reviewer types that form a universally p_s-robust subfamily $\widetilde{\Theta}$ indicated in the theorem, then the sender's ranking over the pooled types is unconditional on p_s. Equivalently, making p_s private does not impact the receiver's correct inference of the reviewer type that represents every pooled cell, which defines the admissible region in the Π-space. Therefore, under an observability structure that falls within two cases in Theorem 5, any equilibrium based on admissibility stays robust even when p_s ceases to be public.

The intuition is simple. In any candidate information structure that the receiver believes to be true, different signals are attached to different values for the sender, creating a ranking of signals in the sender's preference. In case (a), reviewer types can be ordered so that one type sends s^+ on a superset of states, and s^- on the complementary subset, relative to another type. In case (b), different reviewers send different signals constantly independent of the state. Because p_s only reweights states within the simplex, in each of these cases, the sender's preference over reviewers is irrelevant to p_s and depends only on her preference over different signals. In this case, the sender's value function contains full information about the sender's optimal information structure, similarly to direct persuasion under full commitment.

Our examples in the previous section are robust to an unknown p_s because pooling reviewers who send the same signal regardless of the state is exactly a special case of (b) in Theorem 5. In the binary case, robustness fails only when a pool contains both types θ_1 and θ_2 reviewers, who send informative signals but have opposite opinions on each state. If

one of these reviewer types is absent, negligible, or separated by the observability structure, the sender's private prior belief does not directly threaten the equilibrium. As a policy implication, if both θ_1 and θ_2 reviewers are common, a simple solution to this information asymmetry, and a way to maintain equilibrium, is to disclose the proportion of type θ_1 or type θ_2 reviewers, which justifies the policy design to certify faithful reviewers.

This robustness is one-sided. It applies only to the sender's prior belief p_s, but not to the sender's value vector $\mathbf{v}$. The asymmetry comes from the different domains of the two objects. p_s is restricted to the simplex $\Delta(\Omega)$, whereas $\mathbf{v}$ ranges in $\mathbb{R}^n$. Although less likely compared to the sender's prior belief, it is also possible that the receiver is unaware of the sender's objective in persuasion. As with the sender's prior belief, while it is irrelevant in the direct persuasion problem under full commitment, the sender's objective can be important in indicating the receiver's possible deviation if this deviation is not directly observable. However, if the receiver knows p_s but not the sender's objective, then varying $\mathbf{v}$ can potentially reveal any of two reviewer types that induce different signal distributions but are pooled together under that p_s. Accordingly, universal value-vector robustness is much more demanding. If the requirement is imposed for every p_s, full observability is required. In practice, to ensure the robustness of the admissibility region, the policy maker should strive to ensure that at least the sender's objective is publicly known if it is not obvious common knowledge.

In general, this section provides a practical implication for a designer who can refine the observability structure but cannot costlessly reveal every reviewer type. The central implication is not that more disclosure is always better. Rather, the designer should target disclosure to the reviewer categories whose pooling is most likely to threaten admissibility. This approach has been adopted by actual review regulation practice. The current FTC framework does not require the public identification of all reviewers. It explicitly states that firms are not expected to investigate every potential testimonialist. Instead, it targets the categories most likely to create hidden distortions. This practice is particularly important

with a limited budget. Revealing the wrong reviewer types may be expensive while doing little to protect the equilibrium-relevant admissible region. Robustness is therefore not only a theoretical concept, but also provides a discipline for cost-effective design. In that sense, the findings in this section can be read as a guide for how a policy maker should allocate a limited disclosure resource in different scenarios.

8 Concluding Discussion

Canonical Bayesian persuasion takes the commitment to an information structure as a primitive. This simplification isolates the geometry of optimal information. Our paper builds on that foundation and then moves in the complementary direction of opening the implementation black box behind an information structure. By concretely representing information structures as distributions over reviewer mappings, we microfound the commitment assumption as a condition on the observability of the reviewer distribution. This perspective connects the theoretical object of persuasion to empirically observable institutional features. Our insight is therefore not only theoretically interesting but also practically meaningful. For the same reason, the perspective also suggests an empirical route for studying commitment in persuasion through changes in the observability of reviewer composition and their effects on persuasion outcomes.

The finite deterministic reviewer setting in the paper provides a disciplined benchmark and does not imply that real reviewers are always deterministic and non-strategic. In applications, some reviewer types may be stochastic mappings from states to signal distributions, reflecting their backgrounds or institutional incentives. With a finite stochastic basis, the same mechanism continues to apply. The sender's hidden-deviation problem remains a linear program over a polyhedral fiber in the strategy space. What changes is the shape of the fiber and therefore the geometry of admissibility. This is especially relevant for structural empirical work, where the reviewer basis may be estimated rather than imposed. The non-

strategic reviewer in the model should also be read in this reduced-form sense. A strategic intermediary can be integrated into the framework when its incentive constraints generate a stable reporting rule, in which case heterogeneous mappings represent heterogeneous incentive-compatible behavior, including separating behavior across intermediary types. In this sense, the perspective of mapping management applies more broadly whenever persuasion is mediated through agents whose information production can be summarized by reporting technologies. It therefore provides an empirically disciplined test bed for theories in which frictions restrict effective information structures. In those tests, one can ask whether a mechanism survives when the restriction comes from the availability and observability of mappings, rather than being imposed directly on the experiment space.

Our paper also opens natural directions for future work. To isolate the underlying mechanism, we assume that the receiver meets one reviewer, with the meeting probabilities determined directly by the sender's chosen distribution. Future studies may allow the receiver to read multiple reviews, turning our model into a sequential persuasion problem in which the same effective information structure is applied to successive posterior beliefs. In this case, an experiment admissible at the initial belief may become inadmissible after preceding reviews move beliefs to another region, creating a new constraint to shape persuasion. Relaxing the second assumption brings the model closer to word-of-mouth persuasion. A social network filters the sender's reviewer composition into receiver-specific exposure distributions, so the same implementation may generate an admissible experiment for some receivers but not for others. With multiple receivers, this becomes a public persuasion problem with a heterogeneous audience. In both extensions, admissibility becomes asymmetric across histories in the sequential case and across receivers in the network case. This asymmetry may itself be part of the persuasion design. When additional information is unnecessary or counterproductive for some histories or audiences, the limits of commitment can serve as a stabilizing device for persuasion.

References

Antic, Nemanja and Harry Pei, "Selective Disclosure in Overlapping Generations," *arXiv preprint arXiv:2602.09406*, 2026.

Arieli, Itai and Colin Stewart, "Bayesian Persuasion without Commitment," *arXiv preprint arXiv:2511.18662*, 2025.

_ , Yakov Babichenko, and Fedor Sandomirskiy, "Bayesian persuasion with mediators," *arXiv preprint arXiv:2203.04285*, 2022.

Aybas, Yunus C and Eray Turkel, "Persuasion with coarse communication," *arXiv preprint arXiv:1910.13547*, 2025.

Babichenko, Yakov, Inbal Talgam-Cohen, Haifeng Xu, and Konstantin Zabarnyi, "Regret-minimizing Bayesian persuasion," *Games and Economic Behavior*, 2022, *136*, 226–248.

Best, James and Daniel Quigley, "Persuasion for the long run," *Journal of Political Economy*, 2024, *132* (5), 1740–1791.

Bizzotto, Jacopo, Eduardo Perez-Richet, and Adrien Vigier, "Communication via Third Parties," *Working paper*, 2022.

Celik, Levent and Mikhail Drugov, "Score disclosure," *The Economic Journal*, 2025, *135* (666), 519–537.

Chevalier, Judith A and Dina Mayzlin, "The effect of word of mouth on sales: Online book reviews," *Journal of marketing research*, 2006, *43* (3), 345–354.

Corrao, Roberto and Yifan Dai, "The bounds of mediated communication," *arXiv preprint arXiv:2303.06244*, 2024.

Dai, Yifan, Drew Fudenberg, and Harry Pei, "Bayesian Persuasion with Selective Disclosure," *arXiv preprint arXiv:2601.05914*, 2026.

Deb, Rahul, Mallesh M Pai, and Maher Said, "Indirect persuasion," *Journal of Political Economy*, 2026, *134* (4), 1210–1244.

Dworczak, Piotr and Alessandro Pavan, "Preparing for the worst but hoping for the best: Robust (bayesian) persuasion," *Econometrica*, 2022, *90* (5), 2017–2051.

Ederer, Florian and Weicheng Min, "Bayesian Persuasion with Lie Detection," *Working paper*, 2025.

Fudenberg, Drew, Ying Gao, and Harry Pei, "A reputation for honesty," *Journal of Economic Theory*, 2022, *204*, 105508.

Gradwohl, Ronen, Niklas Hahn, Martin Hoefer, and Rann Smorodinsky, "Algorithms for persuasion with limited communication," *Mathematics of Operations Research*, 2022, *47* (3), 2520–2545.

Guo, Yingni and Eran Shmaya, "Costly miscalibration," *Theoretical Economics*, 2021, *16* (2), 477–506.

Hu, Ju and Xi Weng, "Robust persuasion of a privately informed receiver," *Economic Theory*, 2021, *72* (3), 909–953.

Jiang, Shaofei, "Persuasion via Sequentially Acquired Evidence," *Working paper*, 2024.

Kamenica, Emir and Matthew Gentzkow, "Bayesian persuasion," *American Economic Review*, 2011, *101* (6), 2590–2615.

_ **and Xiao Lin**, "Commitment and randomization in communication," *arXiv preprint arXiv:2410.17503*, 2025.

Kosenko, Andrew, "Mediated persuasion," *arXiv preprint arXiv:2012.00098*, 2020.

Kosterina, Svetlana, "Persuasion with unknown beliefs," *Theoretical Economics*, 2022, *17* (3), 1075–1107.

Kreutzkamp, Sophie and Yichuan Lou, "Persuasion without ex-post commitment," *Journal of Economic Theory*, 2025, p. 106058.

Kuvalekar, Aditya, Elliot Lipnowski, and Joao Ramos, "Goodwill in communication," *Journal of Economic Theory*, 2022, *203*, 105467.

Lagziel, David and Ehud Lehrer, "Constrained Mediation: Bayesian Implementability of Joint Posteriors," *arXiv preprint arXiv:2510.20986*, 2025.

Libgober, Jonathan, "False positives and transparency," *American Economic Journal: Microeconomics*, 2022, *14* (2), 478–505.

Lin, Xiao and Ce Liu, "Credible persuasion," *Journal of Political Economy*, 2024, *132* (7), 2228–2273.

Lipnowski, Elliot and Doron Ravid, "Cheap talk with transparent motives," *Econometrica*, 2020, *88* (4), 1631–1660.

Lou, Yichuan, "Private Experimentation, Data Truncation, and Verifiable Disclosure," *arXiv preprint arXiv:2305.04231*, 2023.

Luca, Michael, "Reviews, reputation, and revenue: The case of Yelp.com," *Com (March 15, 2016). Harvard Business School NOM Unit Working Paper*, 2016, (12-016).

_ **and Georgios Zervas**, "Fake it till you make it: Reputation, competition, and Yelp review fraud," *Management science*, 2016, *62* (12), 3412–3427.

Lyu, Qianjun, Wing Suen, and Yimeng Zhang, "Coarse information design," *arXiv preprint arXiv:2305.18020*, 2025.

Margaria, Chiara and Alex Smolin, "Dynamic communication with biased senders," *Games and Economic Behavior*, 2018, *110*, 330–339.

Mathevet, Laurent, David Pearce, and Ennio Stacchetti, "Reputation and information design," *Working Paper*, 2024.

Mayzlin, Dina, Yaniv Dover, and Judith Chevalier, "Promotional reviews: An empirical investigation of online review manipulation," *American Economic Review*, 2014, *104* (8), 2421–2455.

Min, Daehong, "Bayesian persuasion under partial commitment," *Economic Theory*, 2021, *72* (3), 743–764.

Nguyen, Anh and Teck Yong Tan, "Bayesian persuasion with costly messages," *Journal of Economic Theory*, 2021, *193*, 105212.

Pei, Harry, "Repeated communication with private lying costs," *Journal of Economic Theory*, 2023, *210*, 105668.

Perez-Richet, Eduardo and Vasiliki Skreta, "Test design under falsification," *Econometrica*, 2022, *90* (3), 1109–1142.

Rayo, Luis and Ilya Segal, "Optimal information disclosure," *Journal of political Economy*, 2010, *118* (5), 949–987.

Shishkin, Denis, Maria Titova, and Kun Zhang, "Withholding verifiable information," *arXiv preprint arXiv:2206.09918*, 2026.

Titova, Maria and Kun Zhang, "Persuasion with verifiable information," *Journal of Economic Theory*, 2025, p. 106102.

Treust, Maël Le and Tristan Tomala, "Persuasion with limited communication capacity," *Journal of Economic Theory*, 2019, *184*, 104940.

Vong, Allen, "Reputation and efficiency: Information design," *American Economic Journal: Microeconomics*, 2025, *17* (3), 191–243.

Zhou, Junya, "Costly verification and commitment in persuasion," *Journal of Economic Behavior & Organization*, 2023, *212*, 1100–1142.

Appendix

For ease of reference, we use the following notation throughout the appendix.

An observational cell is denoted by $\Theta_k = \{\theta \in \Theta : T_{k\theta} = 1\}$. For a feasible statistic z,

$$\Lambda(z) := \{\lambda \in \Delta(\Theta) : T\lambda = z\}, \qquad \mathcal{P}(z) := \{\Pi(\lambda) : \lambda \in \Lambda(z)\}.$$

Each reviewer type θ induces the deterministic response matrix Π_θ, and

$$\Pi(\lambda) = \sum_{\theta \in \Theta} \lambda(\theta)\Pi_\theta.$$

When a value vector is used with a canonical region $\mathcal{C}_\sigma$, it is written in canonical signal order, so that $\sigma\mathbf{v}$ is the corresponding vector attached to the original signal labels.

Proof of Lemma 1

We prove Lemma 1 by showing that, first, there are n^n deterministic mappings $\Omega \to \mathcal{S}$ when $|\Omega| = |\mathcal{S}| = n$, and second, given these n^n mappings, every column-stochastic matrix $\Pi \in \mathcal{P}$ can be generated by some reviewer distribution λ. In particular, we construct λ such that $\sum_{\theta\in\Theta} \lambda(\theta) = 1$ and $\sum_{\theta\in\Theta} \lambda(\theta)\mathbf{1}\{s = \gamma(\theta, \omega)\} = \pi(s \mid \omega)$ for every $s \in \mathcal{S}$ and $\omega \in \Omega$.

Proof. Let $f : \Omega \to \mathcal{S}$ denote a deterministic mapping, and let θ_f be the reviewer type that implements it, so that $\gamma(\theta_f, \omega) = f(\omega)$ for every $\omega \in \Omega$. Since each of the n states can be assigned to one of n signals, there are n^n such deterministic mappings. Because Θ contains all deterministic mappings and $|\Theta| = n^n$, the correspondence $f \mapsto \theta_f$ is one-to-one.

Fix an arbitrary information structure $\Pi \in \mathcal{P}$. We construct a distribution over deterministic reviewer types. For each deterministic mapping f, define

$$\lambda(\theta_f) := \prod_{\omega\in\Omega} \pi(f(\omega) \mid \omega).$$

This is the probability of drawing the deterministic map f when, for each state ω, a signal is independently drawn according to the column $\pi(\cdot \mid \omega)$ of Π, and the realized state-signal assignments coincide with f.

First, λ is a probability distribution. Indeed,

$$\sum_{\theta \in \Theta} \lambda(\theta) = \sum_{f:\Omega \to \mathcal{S}} \lambda(\theta_f) = \sum_{f:\Omega \to \mathcal{S}} \prod_{\omega \in \Omega} \pi(f(\omega) \mid \omega) = \prod_{\omega \in \Omega} \sum_{s \in \mathcal{S}} \pi(s \mid \omega) = 1,$$

where the third equality expands the product over all choices of one signal for each state, and the last equality follows because each column of Π sums to one.

Second, this distribution over reviewer types induces Π. Fix any $s \in \mathcal{S}$ and $\omega \in \Omega$. Since $f \mapsto \theta_f$ is one-to-one,

$$\sum_{\theta \in \Theta} \lambda(\theta) \mathbf{1}\{s = \gamma(\theta, \omega)\} = \sum_{f:f(\omega)=s} \prod_{\omega' \in \Omega} \pi(f(\omega') \mid \omega').$$

On the right-hand side, the restriction $f(\omega) = s$ fixes the common factor $\pi(s \mid \omega)$. Factoring it out gives

$$\sum_{f:f(\omega)=s} \prod_{\omega' \in \Omega} \pi(f(\omega') \mid \omega') = \pi(s \mid \omega) \prod_{\omega' \neq \omega} \sum_{s' \in \mathcal{S}} \pi(s' \mid \omega') = \pi(s \mid \omega).$$

Thus the effective information structure induced by λ coincides with Π. Since $\Pi \in \mathcal{P}$ was arbitrary, every information structure can be generated by a distribution over deterministic reviewer mappings. □

Proof of Lemma 2

Lemma 2 characterizes the sender's strategy in a pure-strategy equilibrium of the fiber continuation game on $\Lambda(z)$. We show that, conditional on the receiver's action rule, the sender's hidden-deviation incentive is captured by a linear program on a polyhedral constraint, whose optimal solutions assign positive mass only to cellwise maximizing reviewer types.

Proof. Let (λ^*, α^*) be a pure-strategy equilibrium of the fiber continuation game on $\Lambda(z)$. Suppose $\Pi(\lambda^*) \in \mathcal{C}_\sigma$, and let $\mathbf{v}$ be the sender value vector induced by α^*. Conditional on α^*, the sender's payoff from any $\lambda \in \Lambda(z)$ is

$$(\sigma\mathbf{v})^\top \Pi(\lambda) p_s = \sum_{\theta \in \Theta} \lambda(\theta)(\sigma\mathbf{v})^\top \Pi_\theta p_s.$$

The constraint $T\lambda = z$ fixes the total mass in each observational cell. Fix the receiver's action rule and, therefore, the value vector $\mathbf{v}$. Within each cell Θ_k, the sender's problem is to maximize

$$\sum_{\theta \in \Theta_k} \lambda(\theta)(\sigma\mathbf{v})^\top \Pi_\theta p_s$$

subject to

$$\sum_{\theta \in \Theta_k} \lambda(\theta) = z_k, \qquad \lambda(\theta) \geq 0 \quad \forall \theta \in \Theta_k.$$

If $\lambda^*(\theta') > 0$ for some $\theta' \in \Theta_k$ with

$$(\sigma\mathbf{v})^\top \Pi_{\theta'} p_s < \max_{\theta \in \Theta_k} (\sigma\mathbf{v})^\top \Pi_\theta p_s,$$

shifting a small positive amount of mass from θ' to any maximizer $\theta^m \in \Theta_k$ preserves $T\lambda = z$, while strictly increasing the sender's payoff within that cell, and hence her total expected payoff. This is a profitable unilateral deviation from λ^*, contradicting the concept of pure-strategy equilibrium.

Therefore, every reviewer type receiving positive mass in cell k must maximize $(\sigma\mathbf{v})^\top \Pi_\theta p_s$ within that cell. If two types θ' and θ'' both receive positive mass, both must attain the same cellwise maximum. The uniqueness claim follows immediately. □

Proof of Theorem 1

By Definition 3, an information structure Π is purely admissible if and only if there exist a feasible observable statistic z, a reviewer distribution $\lambda^* \in \Lambda(z)$, and an action rule α^*

such that (λ^*, α^*) is a pure-strategy equilibrium of the fiber continuation game and $\Pi = \Pi(\lambda^*) = \sum_{\theta\in\Theta} \lambda^*(\theta)\Pi_\theta$. Theorem 1 connects this equilibrium representation of admissibility to the geometric representation in the theorem. In the necessity direction, we start from an associated pure equilibrium (λ^*, α^*) and construct the cell weights z_k from the equilibrium distribution λ^*. In the sufficiency direction, we start from the geometric representation, construct a reviewer distribution λ^*, and verify that this distribution, together with the receiver's selected action rule, forms a pure-strategy equilibrium in the corresponding fiber.

Proof. Fix $\Pi \in \mathcal{C}_\sigma$, and let $\mathbf{v} = \mathbf{v}(\Pi)$ be the selected value vector induced by the receiver's best response α^Π to Π. For each cell k, define

$$M_k^\sigma(\mathbf{v}) = \arg\max_{\theta\in\Theta_k} (\sigma\mathbf{v})^\top \Pi_\theta p_s.$$

For necessity, suppose that Π is purely admissible, and let (λ^*, α^Π) be an associated pure-strategy equilibrium on some fiber $\Lambda(z)$ such that $\Pi = \Pi(\lambda^*)$. By Lemma 2,

$$\{\theta \in \Theta_k : \lambda^*(\theta) > 0\} \subseteq M_k^\sigma(\mathbf{v}) \qquad \forall k = 1, \ldots, B.$$

Define the cell weights by $z_k := \sum_{\theta\in\Theta_k} \lambda^*(\theta)$. For every cell with $z_k > 0$, define the normalized cell representative

$$\widehat{\Pi}_k := \sum_{\theta\in\Theta_k} \frac{\lambda^*(\theta)}{z_k} \Pi_\theta.$$

If $z_k = 0$, choose any element of

$$\mathrm{co}\{\Pi_\theta : \theta \in M_k^\sigma(\mathbf{v})\},$$

whose value does not affect the sum. Since $\{\theta \in \Theta_k : \lambda^*(\theta) > 0\} \subseteq M_k^\sigma(\mathbf{v})$ for each k,

$$\widehat{\Pi}_k \in \mathrm{co}\{\Pi_\theta : \theta \in M_k^\sigma(\mathbf{v})\}.$$

Grouping the induced information structure by observational cells gives

$$\Pi = \sum_{\theta \in \Theta} \lambda^*(\theta)\Pi_\theta = \sum_{k=1}^{B} z_k \widehat{\Pi}_k.$$

This proves the necessity of the geometric representation.

Conversely, suppose that there exist $z \in \Delta(\{1, \ldots, B\})$ and matrices

$$\widehat{\Pi}_k \in \text{co}\{\Pi_\theta : \theta \in M_k^\sigma(\mathbf{v})\}$$

such that

$$\Pi = \sum_{k=1}^{B} z_k \widehat{\Pi}_k.$$

For each k, choose coefficients $\eta_{\theta k} \geq 0$, supported on $M_k^\sigma(\mathbf{v})$, such that

$$\sum_{\theta \in M_k^\sigma(\mathbf{v})} \eta_{\theta k} = 1, \qquad \widehat{\Pi}_k = \sum_{\theta \in M_k^\sigma(\mathbf{v})} \eta_{\theta k} \Pi_\theta.$$

Construct a reviewer distribution λ^* by setting

$$\lambda^*(\theta) = z_k \eta_{\theta k} \quad \text{for } \theta \in M_k^\sigma(\mathbf{v}),$$

and $\lambda^*(\theta) = 0$ otherwise. By construction,

$$\sum_{\theta \in \Theta_k} \lambda^*(\theta) = z_k \qquad \text{for every } k,$$

so in the reduced representation game, the sender can choose $z = (z_1, \ldots, z_k, \ldots, z_B)$ that defines a fiber $\Lambda(z)$. Moreover,

$$\Pi(\lambda^*) = \sum_{\theta \in \Theta} \lambda^*(\theta)\Pi_\theta = \sum_{k=1}^{B} z_k \widehat{\Pi}_k = \Pi.$$

It remains to verify that (λ^*, α^Π) is a pure-strategy equilibrium of the fiber continuation game on $\Lambda(z)$. The receiver's action rule α^Π is the selected best response to the expected information structure $\Pi = \Pi(\lambda^*)$. For the sender, consider any alternative $\lambda' \in \Lambda(z)$. Since λ' has the same cell masses z_k, and since λ^* assigns mass in each cell only to types in $M_k^\sigma(\mathbf{v})$, the definition of $M_k^\sigma(\mathbf{v})$ implies

$$\sum_{\theta \in \Theta_k} \lambda'(\theta)(\sigma \mathbf{v})^\top \Pi_\theta p_s \leq z_k \max_{\theta \in \Theta_k} (\sigma \mathbf{v})^\top \Pi_\theta p_s = \sum_{\theta \in \Theta_k} \lambda^*(\theta)(\sigma \mathbf{v})^\top \Pi_\theta p_s.$$

Summing over all cells gives

$$(\sigma \mathbf{v})^\top \Pi(\lambda') p_s \leq (\sigma \mathbf{v})^\top \Pi(\lambda^*) p_s.$$

Thus the sender has no profitable within-fiber deviation from λ^*. Therefore (λ^*, α^Π) is a pure-strategy equilibrium of the fiber continuation game, and Π is purely admissible. □

Corollaries to Theorem 1

The analysis of conditions in Theorem 1 generates useful benchmark implications for pure admissibility. Corollary 1 shows how value ties within observational cells enlarge the selected faces and therefore weakly expand the purely admissible region. Corollary 2 studies the two baseline extremes of implementation observability: full observability and complete opacity. Corollary 4 then considers a non-baseline extension in which the observable statistic is a general linear message-distribution statistic.

Proof of Corollary 1

Proof. By Theorem 1, the purely admissible region is obtained by taking convex combinations of the cellwise faces

$$\mathrm{co}\{\Pi_\theta : \theta \in M_k^\sigma(\mathbf{v})\}.$$

If any tie set $M_k^\sigma(\mathbf{v})$ expands, the corresponding convex hull weakly expands. Weighted sums across cells preserve set inclusion, so the purely admissible region weakly expands.

If $M_k^\sigma(\mathbf{v}) = \Theta_k$ for every k, then for any feasible $\Pi = \Pi(\lambda) \in \mathcal{C}_\sigma$, define

$$z_k = \sum_{\theta \in \Theta_k} \lambda(\theta).$$

For $z_k > 0$, set

$$\widehat{\Pi}_k = \sum_{\theta \in \Theta_k} \frac{\lambda(\theta)}{z_k} \Pi_\theta.$$

If $z_k = 0$, choose any $\widehat{\Pi}_k \in \text{co}\{\Pi_\theta : \theta \in \Theta_k\}$. Then

$$\widehat{\Pi}_k \in \text{co}\{\Pi_\theta : \theta \in \Theta_k\}$$

and

$$\Pi = \sum_{k=1}^{B} z_k \widehat{\Pi}_k.$$

Theorem 1 implies that Π is purely admissible. □

Proof of Corollary 2

Proof. If $T\lambda \equiv \lambda$, every observational cell is a singleton. Hence, each fiber is a singleton, and there is no hidden within-fiber deviation. Under the full reviewer space in Lemma 1, every direct information structure can be generated by some reviewer distributions, so every direct-persuasion outcome is purely admissible.

If the observability structure is completely opaque, then there is a single observational cell $\Theta_1 = \Theta$. Applying Theorem 1 gives

$$\Pi \in \text{co}\left\{\Pi_\theta : \theta \in \arg\max_{\theta' \in \Theta} (\sigma \mathbf{v}(\Pi))^\top \Pi_{\theta'} p_s\right\}.$$

Now suppose $p_s \in \text{int}(\Delta(\Omega))$. Let Π be an information structure that induces noncon-

stant posterior beliefs, and suppose its selected value vector has a unique highest component. Let s^* be the signal associated with this unique highest value. For any deterministic reviewer type θ,

$$(\sigma \mathbf{v}(\Pi))^\top \Pi_\theta p_s = \sum_{\omega \in \Omega} p_s(\omega)(\sigma \mathbf{v}(\Pi))_{\gamma(\theta,\omega)},$$

where the subscript of $(\sigma \mathbf{v}(\Pi))_{\gamma(\theta,\omega)}$ denotes the component associated with the signal $\gamma(\theta, \omega)$. Because p_s has full support and s^* is uniquely highest, this expression is uniquely maximized by the constant mapping

$$\gamma(\theta, \omega) = s^* \qquad \forall \omega \in \Omega,$$

which corresponds to an uninformative information structure that induces constant posterior beliefs. In this case, any information structure inducing nonconstant posterior beliefs would have a profitable hidden deviation to this constant mapping and cannot be purely admissible, according to Theorem 1 and Definition 3. Therefore, every purely admissible outcome is equivalent to cheap talk. □

Message-Distribution Observability

The pure-admissibility framework can also reproduce the state-independent result in Lin and Liu (2024) once the observability statistic is allowed to be a general linear statistic. This exercise is useful because it clarifies why the two papers focus on different mechanisms and therefore reach different conclusions.

Corollary 4 (Lin and Liu, 2024)**.** *Fix the sender's prior $p_s \in \Delta(\Omega)$. In the full reviewer space, allow the observability statistic to be a general linear statistic, and define $T^{LL} \in \mathbb{R}^{n \times n^n}$ by*

$$T^{LL}_{k\theta} := \sum_{\omega \in \Omega} p_s(\omega)\pi_\theta(s_k \mid \omega), \qquad k = 1, \ldots, n.$$

Then, for every reviewer distribution $\lambda \in \Delta(\Theta)$,

$$T^{LL}\lambda = \Pi(\lambda)p_s.$$

Thus, the T^{LL}-fiber consists of all implementations that induce the same signal marginal under p_s. If the sender's payoff is state independent, her payoff from any fixed receiver action rule is constant on each T^{LL}-fiber. Message-distribution observability is therefore nonrestrictive for pure admissibility in the state-independent case.

Proof of Corollary 4. For any reviewer distribution $\lambda \in \Delta(\Theta)$, the effective information structure is $\Pi_\lambda := \Pi(\lambda) = \sum_{\theta\in\Theta} \lambda(\theta)\Pi_\theta$. Therefore, for each $s_k \in \mathcal{S} = \{s_1, \ldots, s_n\}$ and $\omega \in \Omega$,

$$\pi_\lambda(s_k \mid \omega) = \sum_{\theta\in\Theta} \lambda(\theta)\pi_\theta(s_k \mid \omega).$$

The probability of signal s_k under this effective information structure and prior p_s is

$$\sum_{\omega\in\Omega} p_s(\omega) \sum_{\theta\in\Theta} \lambda(\theta)\pi_\theta(s_k \mid \omega) = \sum_{\theta\in\Theta} \lambda(\theta) \sum_{\omega\in\Omega} p_s(\omega)\pi_\theta(s_k \mid \omega),$$

which, by the definition of T^{LL}, equals $\sum_{\theta\in\Theta} T^{LL}_{k\theta}\lambda(\theta)$.

Under $\Pi(\lambda)$ and p_s, the k-th component of $T^{LL}\lambda$ is exactly the probability of signal s_k. Therefore,

$$T^{LL}\lambda = \Pi(\lambda)p_s.$$

Since the receiver observes only $T^{LL}\lambda$, the hidden-deviation set generated by this statistic is the T^{LL}-fiber. Now fix the receiver's action rule $a(\cdot) : \mathcal{S} \to \mathcal{A}$. Since the sender's payoff is state independent, her expected payoff under implementation λ, holding $a(\cdot)$ fixed, is

$$\sum_{k=1}^{n} v(a(s_k)) \sum_{\omega\in\Omega} p_s(\omega)\pi_\lambda(s_k \mid \omega) = \sum_{k=1}^{n} v(a(s_k))[T^{LL}\lambda]_k.$$

Therefore, if λ' lies in the same T^{LL}-fiber as λ, then

$$T^{LL}\lambda' = T^{LL}\lambda,$$

and the sender obtains the same expected payoff from λ' and λ, holding the receiver's action

rule fixed. Hence the sender has no profitable hidden deviation within a T^{LL}-fiber. □

Although Corollary 4 technically reproduces Lin and Liu's observability primitive, this representation should not be read as saying that message-distribution observability is a special case of the baseline observability structure in this paper. In our mapping-management model, T is a 0-1 matrix that partitions reviewer mappings into observable cells. The statistic $z = T\lambda$ is therefore a coarse description of the implementation distribution. By contrast, $T^{LL}\lambda$ is the signal marginal under p_s generated by the induced experiment. It is a linear statistic of the implementation distribution, but it is not a partition of reviewer types.

This difference is exactly why the two models have different implications under state-independent sender payoffs. Under implementation observability in our paper, a hidden deviation within a fiber, implemented by substituting reviewer types inside observable cells, may change the payoff-relevant signal marginal. The sender may therefore strictly prefer a different implementation with the same observed statistic if it increases the probability of a higher-value signal. This reflects the mismatch between the observed implementation statistic and the payoff-relevant signal marginal. By contrast, under message-distribution observability, the fiber itself fixes the signal marginal, so the sender's expected payoff from any fixed receiver action rule is unchanged within the fiber.[13] In this case, the statistic $T^{LL}\lambda$ maps every implementation directly into the payoff-relevant message marginal. This alignment neutralizes the mismatch that underlies the commitment problem in our paper. Technically, the T^{LL}-fiber as a feasible deviation set lies inside the sender's indifference set once the receiver's action rule fixes the value vector.

This becomes clearer in the binary case. Write

$$x = \pi(s_2 \mid \omega_1), \qquad y = \pi(s_2 \mid \omega_2), \qquad p = p_s(\omega_1).$$

The probability of signal s_2 under p_s is $m_2 = px + (1-p)y$. Thus, for a fixed observed value

[13] With a common prior, $T^{LL}\lambda$ is exactly the final message distribution and the relevant finding in Lin and Liu (2024) is reproduced. With heterogeneous subjective priors, if the receiver observes a signal marginal under p_r while the sender evaluates payoffs under p_s, the sender may be able to change the payoff-relevant signal marginal under p_s while preserving the receiver's observational statistic under p_r. Then a commitment problem may reappear even with state-independent sender payoffs.

of m_2, a message-distribution fiber is the line

$$px + (1-p)y = m_2.$$

For a fixed receiver action rule, the sender's payoff is

$$(1-m_2)v(a(s_1)) + m_2 v(a(s_2)),$$

which depends on (x, y) only through m_2, and is therefore constant along this line. In the binary square, the message-distribution fiber graphically coincides with the sender's indifference curve. This illustrates why, under state-independent sender payoffs, message-distribution observability makes pure admissibility nonrestrictive.

Proof of Proposition 1

Proposition 1 decomposes the purely admissible region into a finite union of candidate regions. Each candidate region fixes a canonical signal ordering and a candidate set of cellwise sender-maximizing reviewer types, and then applies the geometric condition in Theorem 1. This perspective is more effective than screening the whole strategy space pointwise and will be useful for the binary characterization in Proposition 2.

Therefore, in the proof, we aim to show that any purely admissible point falls within the set defined in the proposition and that any point within this set is purely admissible by the geometric condition in Theorem 1.

Proof. To simplify the expression in the proof, for each $G = (G_1, \ldots, G_B) \in \mathfrak{G}$ and $\sigma \in \Sigma$, define

$$R_{G,\sigma} = \Big\{ \Pi \in \mathcal{C}_\sigma : (\sigma \mathbf{v}(\Pi))^\top \Pi_\theta p_s \geq (\sigma \mathbf{v}(\Pi))^\top \Pi_{\theta'} p_s,\ \forall \theta \in G_k,\ \forall \theta' \in \Theta_k,\ \forall k \Big\}.$$

The expression in the proposition is

$$\bigcup_{G\in\mathfrak{G}}\bigcup_{\sigma\in\Sigma}\left[R_{G,\sigma}\cap \mathrm{co}\{\Pi_\theta : \theta\in\cup_k G_k\}\right].$$

Suppose first that Π is purely admissible. By Theorem 1,

$$\Pi=\sum_{k=1}^{B} z_k\widehat{\Pi}_k,\quad \text{where}\quad \widehat{\Pi}_k\in \mathrm{co}\{\Pi_\theta : \theta\in M_k^\sigma(\mathbf{v}(\Pi))\}.$$

For each k with $z_k>0$, choose a nonempty set $G_k\subseteq M_k^\sigma(\mathbf{v}(\Pi))$ containing the support of some convex representation of $\widehat{\Pi}_k$. For each k with $z_k=0$, choose any nonempty subset $G_k\subseteq M_k^\sigma(\mathbf{v}(\Pi))$. Then $G=(G_1,\ldots,G_B)\in\mathfrak{G}$, $\Pi\in R_{G,\sigma}$, and

$$\Pi\in \mathrm{co}\{\Pi_\theta : \theta\in\cup_k G_k\}.$$

Thus, every purely admissible Π belongs to the union.

Conversely, suppose that Π belongs to one of the sets in the union. Then $\Pi\in\mathcal{C}_\sigma$, and there exist coefficients $\iota_\theta\geq 0$, supported on $\cup_k G_k$, such that $\sum_{\theta\in\cup_k G_k}\iota_\theta=1$ and $\Pi=\sum_{\theta\in\cup_k G_k}\iota_\theta\Pi_\theta$. For each cell, set $z_k:=\sum_{\theta\in G_k}\iota_\theta$. If $z_k>0$, define $\widehat{\Pi}_k:=\sum_{\theta\in G_k}\frac{\iota_\theta}{z_k}\Pi_\theta$. If $z_k=0$, choose any $\widehat{\Pi}_k\in\mathrm{co}\{\Pi_\theta : \theta\in G_k\}$. Then

$$\Pi=\sum_{k=1}^{B} z_k\widehat{\Pi}_k,\quad \text{where}\quad \widehat{\Pi}_k\in \mathrm{co}\{\Pi_\theta : \theta\in G_k\}.$$

Since $\Pi\in R_{G,\sigma}$, every $\theta\in G_k$ is a maximizer in cell k, so

$$G_k\subseteq M_k^\sigma(\mathbf{v}(\Pi)).$$

Therefore

$$\widehat{\Pi}_k\in \mathrm{co}\{\Pi_\theta : \theta\in M_k^\sigma(\mathbf{v}(\Pi))\}.$$

Theorem 1 implies that Π is purely admissible. □

Proof of Proposition 2

Proof. For simplicity, let

$$x := \pi(s_2 \mid \omega_1), \qquad y := \pi(s_2 \mid \omega_2)$$

in the proof. The four reviewer types correspond to the four vertices

$$\rho_{\theta_1} = (0,1), \quad \rho_{\theta_2} = (1,0), \quad \rho_{\theta_3} = (0,0), \quad \rho_{\theta_4} = (1,1).$$

The receiver observes $\lambda(\theta_1)$ and $\lambda(\theta_2)$ separately and observes only $\lambda(\theta_3) + \lambda(\theta_4)$. Hence, the only nontrivial hidden substitution is inside the cell $\{\theta_3, \theta_4\}$.

Suppose first that $x < y$. Then signal s_2 is more likely in state ω_2 than in state ω_1. In the canonical labeling, s_2 is the signal favoring ω_2, so the sender value after s_2 is v_2, while the sender value after s_1 is v_1. Type θ_3 always sends s_1, and type θ_4 always sends s_2. According to Lemma 2, θ_4 is selected within the pooled cell if and only if $v_2 \geq v_1$, and θ_3 is selected if and only if $v_1 \geq v_2$.

If θ_4 is selected, the admissible convex hull is generated by

$$(0,1), \quad (1,0), \quad (1,1),$$

which is the triangle $x + y \geq 1$. If θ_3 is selected, the admissible convex hull is generated by

$$(0,1), \quad (1,0), \quad (0,0),$$

which is the triangle $x + y \leq 1$. Thus when $x < y$, admissibility is equivalent to

$$(v_2 - v_1)(x + y - 1) \geq 0.$$

Now suppose $x > y$. Then signal s_2 favors ω_1, so the value after s_2 is v_1, while the value after s_1 is v_2. Hence θ_4 is selected if and only if $v_1 \geq v_2$, and θ_3 is selected if and only if

$v_2 \geq v_1$. The same triangle argument gives admissibility if and only if

$$(v_2 - v_1)(x + y - 1) \leq 0.$$

On the diagonal $x = y$, the experiment is uninformative, so the two signals induce the same posterior and the same selected sender value. It is therefore purely admissible by the tie argument in Corollary 1. Substituting back $x = \pi(s_2 \mid \omega_1)$ and $y = \pi(s_2 \mid \omega_2)$ gives the result. □

Proof of Theorem 2

In this proof of Theorem 2, we aim to show that first, under conditions in the theorem, the continuation game in the fiber has mixed-strategy equilibria if there is no pure-strategy equilibrium. Second, assuming that the receiver's payoff is affine in the sender's experiment, any mixed-strategy equilibrium can be converted into an equilibrium in which the sender uses the barycenter experiment as a pure strategy.

Proof. Fix a feasible z with nonempty $\mathcal{P}(z)$. Since Θ is finite, $\Delta(\Theta)$ is a simplex. Thus, $\Lambda(z) = \{\lambda \in \Delta(\Theta) : T\lambda = z\}$ is the intersection of this simplex with linear equalities, and is therefore a compact convex polytope. As the affine image of $\Lambda(z)$, $\mathcal{P}(z)$ is also nonempty, compact, and convex.

Consider the mixed extension of the fiber continuation game. The sender's pure strategy space can be taken to be $\mathcal{P}(z)$, and the receiver's pure strategy space is the finite set $\mathcal{A}^{\mathcal{S}}$, given that $\mathcal{A}$ and $\mathcal{S}$ are finite. Since $U(\Pi, \alpha)$ is affine in Π, it is continuous. The sender's payoff from a fixed action rule is also affine in Π. If $\widetilde{\mathbf{v}}(\alpha)$ denotes the value vector attached to the original signal labels, the payoff is $\widetilde{\mathbf{v}}(\alpha)^\top \Pi p_s$. Thus, by the standard existence theorem for compact-continuous games, the mixed extension has a Nash equilibrium.

Let (μ^*, τ^*) be such an equilibrium, where $\mu^* \in \Delta(\mathcal{P}(z))$ is the sender's mixed strategy and $\tau^* \in \Delta(\mathcal{A}^{\mathcal{S}})$ is the receiver's mixed strategy. Starting from this equilibrium, we construct another equilibrium on the same fiber in which the sender uses a pure experiment $\Pi^E \in \mathcal{P}(z)$

and the receiver uses the same mixture τ^*.

Let

$$\Pi^E := \int_{\mathcal{P}(z)} \Pi \, \mu^*(d\Pi)$$

be the barycenter of the sender's equilibrium mixed strategy. Since $\mathcal{P}(z)$ is convex, $\Pi^E \in \mathcal{P}(z)$. We first show that

$$\operatorname{supp} \tau^* \subseteq BR_r(\Pi^E).$$

For any action rule α,

$$\int_{\mathcal{P}(z)} U(\Pi, \alpha) \, \mu^*(d\Pi) = U\left(\int_{\mathcal{P}(z)} \Pi \, \mu^*(d\Pi), \alpha\right) = U(\Pi^E, \alpha),$$

where the first equality uses affineness of U in Π. Hence, a receiver action rule is a best response to the sender's mixed strategy μ^* if and only if it is a best response to the barycenter Π^E. Since τ^* is a receiver best response to μ^*, its support is contained in $BR_r(\Pi^E)$.

Then, we show that Π^E is the sender's best response to the receiver's strategy τ^*. Let σ indicate the canonical region containing Π^E. In the notation of the theorem, $\widetilde{\mathbf{v}}(\alpha) = \sigma \mathbf{v}(\alpha)$, and therefore $\widetilde{\mathbf{v}}(\tau^*) = \sigma \bar{\mathbf{v}}(\tau^*)$. Given τ^*, the sender's payoff from $\Pi \in \mathcal{P}(z)$ is

$$(\sigma \bar{\mathbf{v}}(\tau^*))^\top \Pi p_s.$$

Because μ^* is a sender best response to τ^*, every experiment in the support of μ^* maximizes this affine functional over $\mathcal{P}(z)$. The maximizer set of an affine function over a compact convex set is an exposed face, hence convex. Therefore the barycenter Π^E also belongs to this maximizer face.

Finally, we show that Π^E is sustainable in the indirect persuasion environment. That is, it can be generated by convex combinations of the sender-favored reviewer mappings within each observational cell under the value vector supported by τ^*. Because

$$\mathcal{P}(z) = \sum_{k=1}^{B} z_k \operatorname{co}\{\Pi_\theta : \theta \in \Theta_k\},$$

and that the objective $(\sigma\bar{\mathbf{v}}(\tau^*))^\top \Pi p_s$ is affine in Π, maximizing $(\sigma\bar{\mathbf{v}}(\tau^*))^\top \Pi p_s$ over $\mathcal{P}(z)$ decomposes cell by cell. Within each cell k, the affine objective is maximized on the convex hull of the reviewer types that attain the highest cellwise score. Therefore,

$$\Pi^E \in \sum_{k=1}^{B} z_k \operatorname{co}\Big\{\Pi_\theta : \theta \in \arg\max_{\theta\in\Theta_k} (\sigma\bar{\mathbf{v}}(\tau^*))^\top \Pi_\theta p_s\Big\}.$$

It remains to show that τ^* is nondegenerate. If τ^* were concentrated on a single action rule α, then $\alpha \in BR_r(\Pi^E)$, and Π^E would be a sender best response to α on the fiber. Thus (Π^E, α) would be a pure-strategy equilibrium of the fiber continuation game, contradicting the maintained assumption that no pure-strategy equilibrium exists. Therefore τ^* is nondegenerate. Since $\mathcal{A}^{\mathcal{S}}$ is finite, nondegeneracy of τ^* means $|\operatorname{supp}\tau^*| \geq 2$. Combined with $\operatorname{supp}\tau^* \subseteq BR_r(\Pi^E)$, this verifies that Π^E is supported by receiver mixing on an indifference set, as required by expected admissibility. □

Proof of Theorem 3

Proof. Fix the fiber $\mathcal{P}(z)$, a candidate information structure $\Pi \in \mathcal{P}(z)$, and the receiver's best-response set $BR_r(\Pi)$. For each receiver action rule $\alpha \in BR_r(\Pi)$, the score vector $\mu(\alpha) = \Big((\sigma\mathbf{v}(\alpha))^\top \widetilde{\Pi}_1 p_s, \ldots, (\sigma\mathbf{v}(\alpha))^\top \widetilde{\Pi}_K p_s\Big)$ records the sender's payoff from choosing each extreme point of the fiber, holding the receiver's action rule fixed. A receiver mixture supported on $BR_r(\Pi)$ induces a convex combination of these score vectors. The set of all score vectors that can be generated by receiver mixtures over best responses to Π is

$$C(\Pi) := \operatorname{co}\{\mu(\alpha) : \alpha \in BR_r(\Pi)\}.$$

For a nonempty set $\Phi \subseteq \{1, \ldots, K\}$, the sender is willing to choose any experiment in the face

$$\operatorname{co}\{\widetilde{\Pi}_j : j \in \Phi\}$$

whenever the extreme points in Φ are tied for the highest score and all excluded extreme

points receive weakly lower scores. Fix a nonempty $\Phi \subseteq \{1, \ldots, K\}$ such that $\Pi \in \text{co}\{\widetilde{\Pi}_j : j \in \Phi\}$, which is Condition (IN). Define the corresponding selection cone in score space by

$$D_\Phi = \{x \in \mathbb{R}^K : x_i = x_j \geq x_\ell \text{ for all } i, j \in \Phi,\ \ell \notin \Phi\}.$$

A receiver mixture can make the face Φ sender-optimal if and only if

$$C(\Pi) \cap D_\Phi \neq \varnothing,$$

Since $C(\Pi)$ is compact and convex and D_Φ is a closed convex cone, the separating hyperplane theorem implies that $C(\Pi) \cap D_\Phi \neq \varnothing$ if and only if there is no vector $\beta \in \mathbb{R}^K$ such that

$$\beta^\top x > 0 \ \ \forall x \in C(\Pi), \qquad \beta^\top y \leq 0 \ \ \forall y \in D_\Phi.$$

The second part of this separation condition is exactly what the polar cone records. The polar cone of D_Φ is

$$D_\Phi^\circ = \left\{\beta \in \mathbb{R}^K : \beta^\top \mathbf{1} = 0 \text{ and } \beta_\ell \geq 0 \ \forall \ell \notin \Phi\right\}.$$

Indeed, any $x \in D_\Phi$ can be written as $x = c\mathbf{1} + q$, where $q_i = 0$ for $i \in \Phi$ and $q_\ell \leq 0$ for $\ell \notin \Phi$. The inequality $\beta^\top x \leq 0$ for all such x is equivalent to $\beta^\top \mathbf{1} = 0$ and $\beta_\ell \geq 0$ for every non-selected ℓ. Therefore, $C(\Pi) \cap D_\Phi \neq \varnothing$ if and only if there is no $\beta \in D_\Phi^\circ$ such that

$$\beta^\top x > 0 \qquad \forall x \in C(\Pi).$$

Since $C(\Pi)$ is the convex hull of $\{\mu(\alpha) : \alpha \in BR_r(\Pi)\}$, and $\beta^\top x$ is linear in x, the condition $\beta^\top x > 0$ for every $x \in C(\Pi)$ is equivalent to $\beta^\top \mu(\alpha) > 0$ for every $\alpha \in BR_r(\Pi)$. The above condition is therefore equivalent to

$$\min_{\alpha \in BR_r(\Pi)} \beta^\top \mu(\alpha) \leq 0$$

for every β satisfying

$$\beta^\top \mathbf{1} = 0, \qquad \beta_\ell \geq 0 \quad \forall \ell \notin \Phi,$$

which is exactly condition (SB). Since we have established the equivalence between $C(\Pi) \cap D_\Phi \neq \varnothing$ and Condition (SB), the proof remains to translate this geometric condition back into equilibrium.

Condition (IN) states that Π lies in the convex hull of the selected extreme points. Condition (SB), by the argument above, is equivalent to the existence of a receiver mixture over $BR_r(\Pi)$ whose induced score vector makes those selected extreme points sender-optimal. Therefore, Π is sustained as an equilibrium outcome of the fiber continuation game. Conversely, if Π is sustained as an equilibrium outcome, take the receiver mixture that supports it and let Φ be the set of extreme points that are sender-optimal under the induced score vector and whose convex hull contains Π. Then (IN) holds by construction, and the induced score vector lies in $C(\Pi) \cap D_\Phi$, which is equivalent to (SB).

Under the maintained premise that the fiber has no pure-strategy equilibrium, any receiver mixture that sustains such a Π cannot be degenerate. Otherwise, if the mixture were concentrated on a single action rule $\alpha \in BR_r(\Pi)$, the same argument would make (Π, α) a pure-strategy equilibrium on the fiber. Hence, any Π satisfying (IN) and (SB) is expectedly admissible on the fiber. This completes the proof of Theorem 3. □

Proof of Corollary 3

Proof. Let $m^* := \Pi^* p_s$ be the signal distribution induced by Π^*. By the full-support assumption, the probability of signal s m_s^* is positive for every signal.

Since τ^* is supported on $BR_r(\Pi^*)$, every action rule $\alpha \in \operatorname{supp} \tau^*$ is a receiver best response to Π^*. For each signal s, the sender-favorable value vector $\mathbf{v}(\Pi^*)$ selects a receiver best response that gives the sender the highest value after that signal. Hence,

$$v_s(\alpha) \leq v_s(\Pi^*) \qquad \forall s, \quad \forall \alpha \in \operatorname{supp} \tau^*.$$

Averaging over τ^* gives

$$\bar{v}_s(\tau^*) \leq v_s(\Pi^*) \qquad \forall s.$$

Therefore,

$$(\sigma\bar{\mathbf{v}}(\tau^*))^\top \Pi^* p_s \leq (\sigma\mathbf{v}(\Pi^*))^\top \Pi^* p_s.$$

It remains to show that the inequality is strict. Suppose, toward a contradiction, that equality holds. Since σ only relabels signals and $m^* = \Pi^* p_s$ has full support, the componentwise inequalities above can average to equality only if

$$\bar{v}_s(\tau^*) = v_s(\Pi^*) \qquad \forall s.$$

Since $\bar{v}_s(\tau^*)$ is an average of values weakly below $v_s(\Pi^*)$, this implies that every action rule in $\operatorname{supp}\tau^*$ gives the sender the selected sender-favorable value after every signal:

$$v_s(\alpha) = v_s(\Pi^*) \qquad \forall s, \quad \forall \alpha \in \operatorname{supp}\tau^*.$$

Choose any $\widehat{\alpha} \in \operatorname{supp}\tau^*$. Then by the definition of τ^*, $\widehat{\alpha} \in BR_r(\Pi^*)$ and therefore

$$\mathbf{v}(\widehat{\alpha}) = \mathbf{v}(\Pi^*) = \bar{\mathbf{v}}(\tau^*).$$

Because Π^* is expectedly admissible supported by τ^*, it is a sender best response on its fiber under the expected value vector $\bar{\mathbf{v}}(\tau^*)$. Since $\mathbf{v}(\widehat{\alpha}) = \bar{\mathbf{v}}(\tau^*)$, the same sender-best-response condition holds under the pure receiver action rule $\widehat{\alpha}$. By the concept of equilibrium in a static game, $(\Pi^*, \widehat{\alpha})$ is a pure-strategy equilibrium of the fiber continuation game. Equivalently, Π^* is purely admissible on the fiber. This contradicts the hypothesis that Π^* is not purely admissible by Definition 3. Therefore,

$$(\sigma\bar{\mathbf{v}}(\tau^*))^\top \Pi^* p_s < (\sigma\mathbf{v}(\Pi^*))^\top \Pi^* p_s.$$

□

Proof of Theorem 4

To prove Theorem 4, we construct a class of $n^n-(n-1)^n$ deterministic 0-1 column-stochastic matrices representing different reviewer types, and a class of information structures inside the $n \times n$ column-stochastic experiment space. We then show that (a) every information structure in this class can be represented as a convex combination of these $n^n - (n-1)^n$ vertex matrices, and (b) this class contains a proper fundamental region.

Proof. Fix the designated signal $s_{\hat{i}}$. Let

$$\Theta^{\hat{i}} = \{\theta \in \Theta : \gamma(\theta, \omega) = s_{\hat{i}} \text{ for at least one } \omega \in \Omega\}.$$

When $|\Omega| = |\mathcal{S}| = n$, there are n^n deterministic mappings from Ω to $\mathcal{S}$. The mappings that never use $s_{\hat{i}}$ choose one of the remaining $n-1$ signals in each of the n states. There are therefore $(n-1)^n$ such mappings. These are exactly the mappings excluded from $\Theta^{\hat{i}}$, so we have

$$|\Theta^{\hat{i}}| = n^n - (n-1)^n.$$

Define a class of information structures

$$\mathcal{P}^{\hat{i}} = \left\{ \Pi \in \mathcal{P} : \sum_{j=1}^{n} \pi(s_{\hat{i}} \mid \omega_j) \geq 1 \right\}.$$

It is the intersection of the polytope $\mathcal{P}$ with a closed halfspace, which is therefore a polytope.

We immediately have $\text{co}\{\Pi_\theta : \theta \in \Theta^{\hat{i}}\} \subseteq \mathcal{P}^{\hat{i}}$, since every vertex Π_θ induced by $\theta \in \Theta^{\hat{i}}$ uses $s_{\hat{i}}$ in at least one column and therefore satisfies the defining inequality. For the reverse inclusion $\mathcal{P}^{\hat{i}} \subseteq \text{co}\{\Pi_\theta : \theta \in \Theta^{\hat{i}}\}$, take any $X \in \mathcal{P}^{\hat{i}}$. Let

$$x_j := X_{\hat{i}j}.$$

By definition, $\sum_j x_j \geq 1$. We construct a random deterministic mapping that uses $s_{\hat{i}}$ at least once and whose column marginals equal X.

Choose $Y \sim \text{Unif}[0,1]$. Place intervals $A_j \subset [0,1]$ of lengths $|A_j| = x_j$ consecutively on the unit circle. Since $\sum_j x_j \geq 1$, their union covers $[0,1]$. Let

$$I_j = \mathbf{1}\{Y \in A_j\}.$$

Then $\Pr(I_j = 1) = x_j$, and $\Pr(\exists j : I_j = 1) = 1$.

Define a random deterministic mapping $F : \Omega \to \mathcal{S}$ as follows. If $I_j = 1$, set $F(\omega_j) = s_{\hat{i}}$. If $I_j = 0$, choose $F(\omega_j) = s_i \neq s_{\hat{i}}$ with conditional probability

$$\Pr(F(\omega_j) = s_i \mid I_j = 0) = \frac{X_{ij}}{1 - x_j}.$$

When $x_j = 1$, the event $I_j = 0$ has probability zero, so the conditional rule is irrelevant.

For the designated signal,

$$\Pr(F(\omega_j) = s_{\hat{i}}) = \Pr(I_j = 1) = x_j = X_{\hat{i}j}.$$

For $i \neq \hat{i}$,

$$\Pr(F(\omega_j) = s_i) = \Pr(I_j = 0)\frac{X_{ij}}{1 - x_j} = X_{ij}.$$

Thus, the random deterministic response matrix induced by F, whose (i,j)-entry is $\mathbf{1}\{F(\omega_j) = s_i\}$, has expectation X. Moreover, since $\Pr(\exists j : I_j = 1) = 1$, every realized F uses $s_{\hat{i}}$ at least once. Hence X is a convex combination of deterministic matrices induced by $\Theta^{\hat{i}}$. Since X is defined as an arbitrary element of $\mathcal{P}^{\hat{i}}$, this proves $\mathcal{P}^{\hat{i}} \subseteq \text{co}\{\Pi_\theta : \theta \in \Theta^{\hat{i}}\}$.

Because both $\mathcal{P}^{\hat{i}} \subseteq \text{co}\{\Pi_\theta : \theta \in \Theta^{\hat{i}}\}$ and $\text{co}\{\Pi_\theta : \theta \in \Theta^{\hat{i}}\} \subseteq \mathcal{P}^{\hat{i}}$ are true, we have

$$\mathcal{P}^{\hat{i}} = \text{co}\{\Pi_\theta : \theta \in \Theta^{\hat{i}}\}.$$

Since the right-hand side is the convex hull of deterministic vertices that are also vertices of $\mathcal{P}$, the vertices of $\mathcal{P}^{\hat{i}}$ are exactly the deterministic response matrices induced by $\Theta^{\hat{i}}$.

It remains to show that a proper fundamental region lies inside $\mathcal{P}^{\hat{i}}$. For any $\Pi \in \mathcal{P}$, the

row sums satisfy

$$\sum_{i=1}^{n}\sum_{j=1}^{n}\pi(s_i \mid \omega_j) = n.$$

Therefore, some row has sum at least 1. Relabel signals so that this row becomes $s_{\hat{i}}$. The relabeled experiment belongs to $\mathcal{P}^{\hat{i}}$. Selecting one representative from each relabeling class, using a fixed rule when ties arise, gives a proper fundamental region $\mathcal{D} \subset \mathcal{P}^{\hat{i}}$. Since every element of $\mathcal{P}^{\hat{i}}$ is implementable using $\Theta^{\hat{i}}$, every $\Pi \in \mathcal{D}$ is implementable using only reviewer types in $\Theta^{\hat{i}}$. □

Proof of Proposition 3

Proof. As in the proof of Proposition 2, let

$$x = \pi(s_2 \mid \omega_1), \qquad y = \pi(s_2 \mid \omega_2).$$

The four deterministic reviewer types are the vertices

$$\rho_{\theta_1} = (0,1), \quad \rho_{\theta_2} = (1,0), \quad \rho_{\theta_3} = (0,0), \quad \rho_{\theta_4} = (1,1).$$

First, suppose that the designer knows a labeled representative $\Pi^\dagger = (x, y)$ of the target outcome. A three-cell observability structure pools one pair of vertices and leaves the other two vertices as singleton cells. If the pooled pair consists of two adjacent vertices, the robust admissible region is the intersection of the two triangles obtained by selecting either member of the pooled pair. The four adjacent pooled pairs generate following four regions

$$\{y \geq x,\ x+y \geq 1\}, \quad \{y \geq x,\ x+y \leq 1\}, \quad \{y \leq x,\ x+y \geq 1\}, \quad \{y \leq x,\ x+y \leq 1\}.$$

These four regions cover the square as the full space of the binary case. Hence, for any labeled target (x, y), the designer can choose a pooled adjacent pair whose robust admissible region contains (x, y). Then $\Pi^\dagger$ is purely admissible.

Now suppose that the labeled target is not known, but the relevant value ranking is

known. Consider the case $v_1 \geq v_2$, which is symmetric to the other case. Choose the three-cell structure that pools θ_2 with θ_3 and leaves θ_1 and θ_4 as singleton cells. If $x < y$, signal s_2 favors ω_2, so the value after s_1 is v_1 and the value after s_2 is v_2. In the pooled cell, θ_3 is weakly preferred to θ_2. The selected hull is

$$\text{co}\{(0,0),(0,1),(1,1)\},$$

which is the region $y \geq x$. Thus the entire canonical region $x < y$ is admissible. Since every persuasion outcome has a representative in this canonical region after signal relabeling, the target outcome is sustained up to relabeling. The case $v_2 \geq v_1$ follows by the symmetric construction. □

Proof of Theorem 5

The disagreement set characterizing states where two reviewer types send different signals is the central building block for the proof of Theorem 5. We first show in a pairwise lemma that the sender's preference between two reviewer types is robust to variation in her prior belief if and only if the ordered pair of signals sent by these two types is the same at every disagreement state.[14] Based on this lemma, we use two reviewer types with a largest disagreement set to characterize the reviewer subfamilies that ensure p_s-robustness.

We first prove the key lemma. For $\theta', \theta'' \in \Theta$, define

$$D(\theta', \theta'') = \{\omega \in \Omega : \gamma(\theta', \omega) \neq \gamma(\theta'', \omega)\}.$$

Lemma A.1. *For any two types θ', θ'', the expression*

$$\mathbf{v}^\top (\Pi_{\theta'} - \Pi_{\theta''}) p_s$$

[14]Here, "ordered" refers to which type sends which signal. For example, (a, b) means that on the same state, the first type sends a and the second type sends b. Therefore, (a, b) and (b, a) are different ordered pairs.

does not change sign as p_s varies over $\Delta(\Omega)$, for every fixed signal-value vector $\mathbf{v}$, if and only if all states in $D(\theta', \theta'')$ carry the same ordered signal pair. Equivalently, either $D(\theta', \theta'') = \varnothing$, or there exist distinct signals $a, b \in \mathcal{S}$ such that

$$\gamma(\theta', \omega) = a, \qquad \gamma(\theta'', \omega) = b \quad \forall \omega \in D(\theta', \theta'').$$

Proof of Lemma A.1. If $D(\theta', \theta'') = \varnothing$, then $\Pi_{\theta'} = \Pi_{\theta''}$, and the expression is identically zero. Hence suppose $D(\theta', \theta'') \neq \varnothing$. With this supposition, if the structure in Lemma A.1 holds, then

$$\mathbf{v}^\top (\Pi_{\theta'} - \Pi_{\theta''}) p_s = (v_a - v_b) \sum_{\omega \in D(\theta', \theta'')} p_s(\omega).$$

The sign of this comparison is determined by $(v_a - v_b)$, not by $\sum_{\omega \in D(\theta', \theta'')} p_s(\omega)$. Hence, this sign cannot reverse as p_s varies.

The converse can be proved by contradiction. If there are two disagreement states ω_0, ω_1 with different ordered signal pairs

$$(\gamma(\theta', \omega_0), \gamma(\theta'', \omega_0)) = (a, b), \qquad (\gamma(\theta', \omega_1), \gamma(\theta'', \omega_1)) = (c, d),$$

where $(a, b) \neq (c, d)$, then for certain value vectors, such as $\mathbf{v} = (e_a - e_b) - (e_c - e_d)$, where e_s is the standard unit vector corresponding to signal s, we have

$$v_a - v_b > 0, \qquad v_c - v_d < 0.$$

The strict inequalities follow because $(a, b) \neq (c, d)$ implies $e_a - e_b \neq e_c - e_d$.

In this case, putting all prior mass on ω_0 or on ω_1 reverses the sign, which creates a contradiction. Therefore, all disagreement states have the same ordered signal pair. This proves Lemma A.1. □

Proof of Theorem 5. Now suppose $\widetilde{\Theta}$ satisfies universal p_s-robustness. If all types in $\widetilde{\Theta}$ are identical, take $\widetilde{\Omega} = \varnothing$, and the result is immediate. Otherwise, choose $\underline{\theta}, \overline{\theta} \in \widetilde{\Theta}$ such that $|D(\underline{\theta}, \overline{\theta})|$ is maximal. Define $\widetilde{\Omega} := D(\underline{\theta}, \overline{\theta})$. By Lemma A.1, after naming two signals s^- and s^+, we may write

$$\gamma(\underline{\theta}, \omega) = s^-, \qquad \gamma(\overline{\theta}, \omega) = s^+ \quad \forall \omega \in \widetilde{\Omega},$$

and

$$\bar{\gamma}(\omega) = \gamma(\underline{\theta}, \omega) = \gamma(\overline{\theta}, \omega) \quad \forall \omega \in \Omega \setminus \widetilde{\Omega}.$$

We first claim that every type $\theta \in \widetilde{\Theta}$ agrees with $\bar{\gamma}$ on $\Omega \setminus \widetilde{\Omega}$. Suppose not. Then for some state $x \notin \widetilde{\Omega}$,

$$\gamma(\theta, x) \neq \bar{\gamma}(x).$$

Since both $\underline{\theta}$ and $\overline{\theta}$ send $\bar{\gamma}(x)$ at x, the ordered disagreement pair between θ and each of these two types at x is $(\gamma(\theta, x), \bar{\gamma}(x))$.

Apply Lemma A.1 to $(\theta, \underline{\theta})$. On $\widetilde{\Omega}$, $\underline{\theta}$ always sends s^-. Therefore, if θ disagrees with $\underline{\theta}$ at any state in $\widetilde{\Omega}$, the ordered signal pair there must also be $(\gamma(\theta, x), \bar{\gamma}(x))$, which is possible only if $\bar{\gamma}(x) = s^-$. If $\bar{\gamma}(x) \neq s^-$, θ must agree with $\underline{\theta}$ on all of $\widetilde{\Omega}$. Similarly, apply Lemma A.1 to $(\theta, \overline{\theta})$. Because $\overline{\theta}$ always sends s^+ on $\widetilde{\Omega}$, if $\bar{\gamma}(x) \neq s^+$, then θ must agree with $\overline{\theta}$ on all of $\widetilde{\Omega}$. Since $s^- \neq s^+$, the signal $\bar{\gamma}(x)$ cannot be equal to both. If $\bar{\gamma}(x) \notin \{s^-, s^+\}$, then θ would have to agree with both $\underline{\theta}$ and $\overline{\theta}$ on all of $\widetilde{\Omega}$, which is impossible because these two types disagree on every state in $\widetilde{\Omega}$. This impossibility implies $\bar{\gamma}(x) \in \{s^-, s^+\}$. Equivalently, θ must agree with $\underline{\theta}$ or $\overline{\theta}$, but not both, on all $\widetilde{\Omega}$.

However, $\bar{\gamma}(x) \in \{s^-, s^+\}$ is also impossible under the definition of $|D(\underline{\theta}, \overline{\theta})|$. If $\bar{\gamma}(x) = s^-$ so that $\bar{\gamma}(x) \neq s^+$, θ must agree with $\overline{\theta}$ on all of $\widetilde{\Omega}$. That is, θ disagrees with $\underline{\theta}$ not only on all of $\widetilde{\Omega}$, but also at $x \in \Omega \setminus \widetilde{\Omega}$. This contradicts the definition that $|D(\underline{\theta}, \overline{\theta})|$ is maximal because now $|D(\underline{\theta}, \theta)| > |D(\underline{\theta}, \overline{\theta})|$. The case $\bar{\gamma}(x) = s^+$ is symmetric. Because of this contradiction, there is no x such that $\gamma(\theta, x) \neq \bar{\gamma}(x)$, which proves our first claim that every type in $\widetilde{\Theta}$ agrees with $\bar{\gamma}$ outside $\widetilde{\Omega}$.

Fix any $\theta \in \widetilde{\Theta}$. Since all types agree with $\bar{\gamma}$ outside $\widetilde{\Omega}$, all disagreements between θ and the two extreme types occur inside $\widetilde{\Omega}$. Compare θ with $\underline{\theta}$. Since $\underline{\theta}$ sends s^- on every state in

$\widetilde{\Omega}$, Lemma A.1 implies that whenever θ differs from $\underline{\theta}$, θ must send the same signal across all such disagreement states. Call this signal c_θ with $c_\theta = s^-$ if there is no such disagreement. We therefore have

$$\gamma(\theta, \omega) \in \{s^-, c_\theta\} \qquad \forall \omega \in \widetilde{\Omega}.$$

Similarly, comparing θ with $\overline{\theta}$, and using that $\overline{\theta}$ sends s^+ on every state in $\widetilde{\Omega}$, there is a signal d_θ, with $d_\theta = s^+$ if there is no such disagreement, such that

$$\gamma(\theta, \omega) \in \{s^+, d_\theta\} \qquad \forall \omega \in \widetilde{\Omega}.$$

Thus the set of signals used by θ on $\widetilde{\Omega}$ is contained in both $\{s^-, c_\theta\}$ and $\{s^+, d_\theta\}$. If θ uses only one signal on $\widetilde{\Omega}$, then it is constant on $\widetilde{\Omega}$. If it uses two signals, then the two sets above must coincide. Since one contains s^- and the other contains s^+, this can happen only when the two signals are exactly s^- and s^+. Therefore every type either is constant on $\widetilde{\Omega}$, or uses only the two signals s^- and s^+ on $\widetilde{\Omega}$.

If some type is constant on $\widetilde{\Omega}$ at a signal not in $\{s^+, s^-\}$, then every type must be constant on $\widetilde{\Omega}$. If not every type is constant on $\widetilde{\Omega}$, then by the preceding classification there must be a type that uses both s^+ and s^-. Comparing that constant type with a type using both s^+ and s^- would produce two different ordered disagreement pairs, one where the latter sends s^+, and another where it sends s^- on states in $\widetilde{\Omega}$, which contradicts Lemma A.1. Hence, case (b) in the theorem holds.

In the other case, all types use only s^+ and s^- on $\widetilde{\Omega}$. Define

$$\Omega_\theta = \{\omega \in \widetilde{\Omega} : \gamma(\theta, \omega) = s^+\}.$$

If two sets $\Omega_{\theta'}$ and $\Omega_{\theta''}$ crossed, there would exist

$$\omega \in \Omega_{\theta'} \setminus \Omega_{\theta''}, \qquad \omega' \in \Omega_{\theta''} \setminus \Omega_{\theta'}.$$

Then

$$(\gamma(\theta', \omega), \gamma(\theta'', \omega)) = (s^+, s^-),$$

whereas

$$(\gamma(\theta', \omega'), \gamma(\theta'', \omega')) = (s^-, s^+),$$

contradicting Lemma A.1. Thus, the family $\{\Omega_\theta\}_{\theta \in \widetilde{\Theta}}$ is totally ordered by inclusion, and case (a) holds.

The converse is immediate. If case (a) holds, then for any two types, nestedness implies that all disagreement states have the same ordered pair, either (s^+, s^-) or (s^-, s^+). Therefore

$$\mathbf{v}^\top (\Pi_{\theta'} - \Pi_{\theta''}) p_s$$

is equal to $\pm(v_{s^+} - v_{s^-}) \sum_{\omega \in D(\theta', \theta'')} p_s(\omega)$. If case (b) holds, any two types differ only by constant signals $(s_{\theta'}, s_{\theta''})$ on $\widetilde{\Omega}$, so the same expression is $(v_{s_{\theta'}} - v_{s_{\theta''}}) \sum_{\omega \in \widetilde{\Omega}} p_s(\omega)$. In both cases, p_s only changes a weakly nonnegative scalar multiplying a fixed value difference, so the sign cannot reverse with p_s. This proves universal p_s-robustness. □